

Credit Risk Modeling Using SAS[®]

Course Notes

Teacher : Christophe Mues

C.MUES@SOTON.AC.UK

Credit Risk Modeling Using SAS® Course Notes was developed by Bart Baesens. Editing and production support was provided by the Curriculum Development and Support Department.

SAS and all other SAS Institute Inc. product or service names are registered trademarks or trademarks of SAS Institute Inc. in the USA and other countries. ® indicates USA registration. Other brand and product names are trademarks of their respective companies.

Credit Risk Modeling Using SAS® Course Notes

Copyright © 2011 Bart Baesens. All rights reserved. Printed in the United States of America. No part of this publication may be reproduced, stored in a retrieval system, or transmitted, in any form or by any means, electronic, mechanical, photocopying, or otherwise, without the prior written permission of the publisher, SAS Institute Inc.

Book code E2054, course code BB4C71, prepared date 31Aug2011.

BB4C71_001

ISBN 978-1-61290-083-4

Table of Contents

Course Description	vi
Prerequisites	vii
Chapter 1 Introduction	1-1
1.1 Introduction.....	1-3
Chapter 2 An Introduction to Credit Scoring.....	2-1
2.1 Introduction to Credit Scoring	2-3
Chapter 3 Basel I, Basel II, and Basel III	3-1
3.1 Basel I, Basel II, and Basel III	3-3
Chapter 4 Preprocessing Data for Credit Scoring and PD Modeling	4-1
4.1 Preprocessing Data.....	4-3
Chapter 5 Classification Techniques for Credit Scoring and PD Modeling.....	5-1
5.1 Classification Techniques	5-3
Chapter 6 Measuring the Performance of Credit Scoring Classification Models	6-1
6.1 Measuring Performance	6-3
Chapter 7 Setting the Classification Cutoff	7-1
7.1 Setting the Classification Cutoff.....	7-3
Chapter 8 Input Selection for Classification.....	8-1
8.1 Input Selection	8-3

Chapter 9 Implementing Scorecards.....	9-1
9.1 Implementing Scorecards.....	9-3
Chapter 10 Reject Inference.....	10-1
10.1 Reject Inference	10-3
Chapter 11 Behavioral Scoring.....	11-1
11.1 Behavioral Scoring	11-3
Chapter 12 Defining Default Ratings and Calibrating PD.....	12-1
12.1 Defining Default Ratings and Calibrating PD	12-3
Chapter 13 LGD Modeling	13-1
13.1 LGD Modeling.....	13-3
Chapter 14 Validation of Basel II Models	14-1
14.1 Validating Basel II Models.....	14-3
Chapter 15 Low Default Portfolios	15-1
15.1 Low Default Portfolios	15-3
Chapter 16 Stress Testing	16-1
16.1 Stress Testing	16-3
Chapter 17 New Techniques for PD/LGD Modeling: Neural Networks.....	17-1
17.1 Neural Networks	17-3
Chapter 18 New Techniques for PD/LGD Modeling: Support Vector Machines	18-1
18.1 Support Vector Machines	18-3

Chapter 19	New Techniques for PD/LGD Modeling: Survival Analysis	19-1
19.1	Survival Analysis	19-3
Appendix A	Exercises	A-1
A.1	Exercises	A-3
Appendix B	Data Dictionary for the Applicants Data Set.....	B-1
B.1	Data Dictionary	B-3
Appendix C	References	C-1
C.1	References.....	C-3

Course Description

In this course, students learn how to develop credit risk models in the context of the recent Basel II and Basel III guidelines. The course provides a sound mix of both theoretical and technical insights, as well as practical implementation details. These are illustrated by several real-life case studies and exercises.

To learn more...



For information on other courses in the curriculum, contact the SAS Education Division at 1-800-333-7660, or send e-mail to training@sas.com. You can also find this information on the Web at support.sas.com/training/ as well as in the Training Course Catalog.



For a list of other SAS books that relate to the topics covered in this Course Notes, USA customers can contact our SAS Publishing Department at 1-800-727-3228 or send e-mail to sasbook@sas.com. Customers outside the USA, please contact your local SAS office.

Also, see the Publications Catalog on the Web at support.sas.com/pubs for a complete list of books and a convenient order form.

Prerequisites

Before attending this course, you should have business expertise in credit risk and a basic understanding of statistical classification methods. Previous SAS software and SAS Enterprise Miner experience is helpful but not necessary. Contact instructor Bart Baesens directly if you have questions:
Bart.Baesens@econ.kuleuven.be.

Chapter 1 Introduction

1.1 Introduction.....1-3

1.1 Introduction

Lecturer: Bart Baesens

- Studied at the Catholic University of Leuven (Belgium)
 - Business Engineer in Management Informatics, 1998
 - Ph.D. in Applied Economic Sciences, 2003
- Ph.D. Title: Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques
- Associate professor at the K.U.Leuven, Belgium
- Associate professor at Vlerick Leuven Ghent Management School
- Lecturer at the School of Management at the University of Southampton, United Kingdom
- Research: data mining, neural networks, SVMs, rule extraction, survival analysis, CRM, credit scoring, Basel II, ...
- Twitter: DataMiningApps
- Facebook: Data Mining with Bart
- www.dataminingapps.com



3

Lecturer: Christophe Mues

- Studied at the Catholic University of Leuven (Belgium)
 - Business Engineer in Management Informatics, 1994
 - Ph.D. in Applied Economic Sciences, 2002
- Ph.D. Title: On the Use of Decision Tables and Diagrams in Knowledge Modeling and Verification
- Senior lecturer at the School of Management of the University of Southampton, United Kingdom
- Continuing research collaboration with K.U.Leuven, Belgium
- Research: business intelligence, data mining, decision tables, credit scoring, Basel II



4

Software Support

- SAS/STAT
 - PROC MEANS, PROC FREQ,
PROC LOGISTIC, PROC IML, ...
- SAS/INSIGHT
 - interactive data analysis
- SAS Enterprise Miner
- SAS Credit Scoring Nodes
 - interactive grouping, scorecard,
reject inference, credit exchange

5

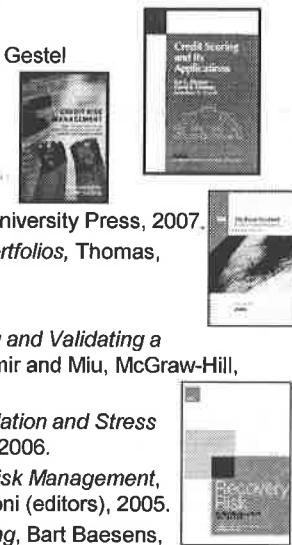
Day Schedule

9:00–10:30	Morning Session 1
10:30–10:45	Morning Break
10:45–12:30	Morning Session 2
12:30–1:30	Lunch
1:30–3:15	Afternoon Session 1
3:15–3:30	Afternoon Break
3:30–5:00	Afternoon Session 2
5:00	End of Day

6

Relevant Background: Books

- *Credit Risk Management: Basic Concepts*, Van Gestel and Baesens, Oxford University Press, 2008.
- *Credit Scoring and Its Applications*, Thomas, Edelman, and Crook. Siam Monographs on Mathematical Modeling and Computation, 2002.
- *The Credit Scoring Toolkit*, Anderson, Oxford University Press, 2007.
- *Consumer Credit Models: Pricing, Profit and Portfolios*, Thomas, Oxford University Press, 2009.
- *The Basel Handbook*, Ong, 2004.
- *Basel II Implementation: A Guide to Developing and Validating a Compliant Internal Risk Rating System*, Ozdemir and Miu, McGraw-Hill, 2008.
- *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, Engelmann and Rauhmeier, Springer, 2006.
- *Recovery Risk: The Next Challenge in Credit Risk Management*, Edward Altman, Andrea Resti, and Andrea Sironi (editors), 2005.
- *Developing Intelligent Systems for Credit Scoring*, Bart Baesens, Ph.D. thesis, K.U.Leuven, 2003.



7

Relevant Background: Journals

- *Journal of Credit Risk*
- *Risk Magazine*
- *Journal of Banking and Finance*
- *Journal of the Operational Research Society*
- *European Journal of Operational Research*
- *Management Science*
- *IMA Journal of Mathematics*

8

Background Papers (Bart + Christophe)

- VAN GESTEL T., MARTENS D., VANDEN BRANDEN K., BAESENS B., A Practical Framework for Credit Risk Stress Testing, *The Journal of Risk Model Validation*, forthcoming, 2011.
- LOTERMAN G., BROWN I., MARTENS D., MUES C., BAESENS B., Benchmarking Regression Algorithms for Loss Given Default Modeling, *International Journal of Forecasting*, forthcoming 2011.
- VAN GOOL J., VERBEKE W., SERCU P., BAESENS B., Credit Scoring for Microfinance – is it worth it?, *International Journal of Finance and Economics*, forthcoming, 2011.
- VAN GESTEL T., MARTENS D., BAESENS B., From Linear to Non-linear Kernel Based Classifiers for Bankruptcy Prediction, *Neurocomputing*, Volume 73, Number 16–18, pp. 2955–2970, 2010.
- VAN LAERE E., BAESENS B., The development of a simple and intuitive rating system under Solvency II, *Insurance: Mathematics and Economics*, Volume 46, Issue 3, pp. 500–510, 2010.
- CASTERMANS G., MARTENS D., VAN GESTEL T., HAMERS B., BAESENS B., An Overview and Framework for PD Backtesting and Benchmarking, *Journal of the Operational Research Society*, 61, pp. 359–373, 2010.
- VAN GESTEL T., MARTENS D., BAESENS B., FEREMANS D., HUYSMANS J., VANTHIENEN J., Forecasting and Analyzing Insurance Companies' Ratings, *International Journal of Forecasting*, 23 (3), pp. 513–529, 2007.
- VAN GESTEL T., BAESENS B., VAN DIJCKE P., SUYKENS J., GARCIA J. AND ALDERWEIRELD T., Linear and nonlinear credit scoring by combining logistic regression and support vector machines, *Journal of Credit Risk*, Volume 1, Number 4, 2005.
- VAN GESTEL T., BAESENS B., VAN DIJCKE P., GARCIA J., SUYKENS J.A.K., VANTHIENEN J., A process model to develop an internal rating system: sovereign credit ratings, *Decision Support Systems*, 42 (2), pp. 1131–1151, 2006.
- BAESENS B., SETIONO R., MUES C., VANTHIENEN J., Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation, *Management Science*, 49 (3), pp. 312–329, March 2003.
- BAESENS B., VAN GESTEL T., VIAENE S., STEPANOVA M., SUYKENS J., VANTHIENEN J., Benchmarking State of the Art Classification Algorithms for Credit Scoring, *Journal of the Operational Research Society*, 54 (6), pp. 627–635, 2003.

Relevant Background: Web Sites

www.defaultrisk.com

<http://www.bis.org/>

Web sites of regulators:

- www.fsa.gov.uk (United Kingdom)
- www.hkma.gov.uk (Hong Kong)
- www.apra.gov.au (Australia)
- www.mas.gov.sg (Singapore)

In the U.S.:

- Four agencies: OCC, Federal Reserve, FDIC, OTS
- Proposed Supervisory Guidance for Internal Ratings Based Systems for Credit Risk, Advanced Measurement Approaches for Operational Risk, and the Supervisory Review Process (Pillar 2), **Federal Register**, Vol. 72, No. 39, February 2007.
- Risk-based capital standards: Advanced Capital Adequacy Framework-Basel II; Final rule, **Federal Register**, December 2007.

Relevant Background: Conferences and Workshops

- Edinburgh conference on Credit Scoring and Credit Control
 - Next one in 2013
- Credit Scoring workshops in Belgium/Southampton
- SAS courses
- Data mining conferences
 - KDD, PKDD/ECML, PAKDD, SAS Analytics conferences

11

Course Overview

- Credit Scoring
- Basel I, Basel II and Basel III
- Data preprocessing for PD modeling
- Classification techniques for PD modeling
- Measuring performance of PD models
- Specific PD modeling issues
- Defining ratings and PD calibration
- LGD modeling and EAD modeling
- Validation of Basel II IRB models
- Low default portfolios
- Stress testing
- New techniques for PD/LGD/EAD modeling
 - Neural networks (short)
 - SVMs (short)
 - Survival analysis

Only in four-day version!

12

Chapter 2 An Introduction to Credit Scoring

2.1 Introduction to Credit Scoring	2-3
---	------------

2.1 Introduction to Credit Scoring

Credit Scoring

- Estimate whether applicant will successfully repay his/her loan based on various information
 - Applicant characteristics (for example, age, income, employment status, time at address, ...)
 - Credit bureau information
 - Application information of other applicants
 - Repayment behavior of other applicants
- Develop models (also called *scorecards*) estimating the probability of default of a customer
- Typically, assign points to each piece of information, add all points, and compare with a threshold (cutoff)

3

Judgmental versus Statistical Approach

Judgmental

- Based on experience
- Five C's: character, capital, collateral, capacity, condition

Statistical

- Based on multivariate correlations between inputs and risk of default

Both assume that the future will resemble the past.

4

Benefits of Developing Credit Scoring Models

- Speed and accuracy
- Consistency
- Reduced bad debt loss
- Reduced operating costs
- Improved portfolio management

5

Types of Credit Scoring

- Application scoring
- Behavioral scoring
- Dynamic scoring
- Bankruptcy prediction

6

Application Scoring

- Estimate probability of default at the time the applicant applies for the loan!
- Use a predetermined definition of default.
 - For example, three months of payment arrears
 - Has now been fixed in Basel II
- Use application variables (age, income, marital status, years at address, years with employer, known client, ...).

7

continued...

Application Scoring

- Use bureau variables.
 - Bureau score versus raw bureau data (for example, number of credit checks, total amount of credits, delinquency history, ...)
 - In the U.S.:
 - FICO scores: between 300 to 850
 - Experian, Equifax, TransUnion
 - Others:
 - Baycorp Advantage (Australia and New Zealand)
 - Schufa (Germany)
 - BKR (the Netherlands)
 - CKP (Belgium)
 - Dun & Bradstreet (MidCorp, SME)

8

Example Application Scorecard

Let cutoff = 500

So, a new customer applies for credit:

AGE	32	120 points
GENDER	Female	180 points
SALARY	\$1.150	160 points

Total	460 points
--------------	-------------------

REFUSE CREDIT

Characteristic Name	Attribute	Scorecard Points
AGE 1	Up to 26	100
AGE 2	26 - 35	120
AGE 3	35 - 37	185
AGE 4	37+	225
GENDER 1	Male	90
GENDER 2	Female	180
SALARY 1	Up to 500	120
SALARY 2	501-1000	140
SALARY 3	1001-1500	160
SALARY 4	1501-2000	200
SALARY 5	2001+	240

9

...

Application Scoring

Snapshot 1

Application data

Age
Income
Marital status
Savings amount
....

Snapshot 2

t_{18}

Good or Bad Payer?

Credit Bureau data

Bureau score
Delinquency history
Number of bureau checks
Number of outstanding credits
....

10

Example Variables Used in Application Scorecard Development

Age	Income	Number of trades R1, R2...R9
Time at residence	Total Liabilities	% trades R1-R2
Time at employment	Total Debt	% Trades R3 or worse
Time in industry	Total Debt service \$	% trades O1..O9
First digit of postal code	Total Debt Service Ratio	% trades O1-O2
Geographical: Urban/Rural/Regional/Provincial	Gross Debt Service Ratio	% Trades O3 or worse
Residential Status	Revolving Debt/Total Debt	Total credit lines - revolving
Employment status	Other Bank card	Total balance - revolving
Lifestyle code	Number of other bank cards	Total Utilization
Existing client (Y/N)	Total Trades	total balance- all
Years as client	Trades opened in last six months	
Number of products internally (internal bankruptcy/write-off)	Trades Opened in Last six months / Total Trades	
Number of delinquencies	Total Inactive Trades	
Beacon/Empirical	Total Revolving trades	
Time at bureau	Total term trades	
Total Inquiries	% revolving trades	
Time since last inquiry	Total term trades	
Inquiries in the last 3/6/12 mths	Number of trades current	
Inquiries in the last 3/6/12 mths as % of total		

11

Application Scoring: Summary

- New customers applying for credit
- Two snapshots of the state of the customer
 - Second snapshot typically taken around 18 months after loan origination
- Application scorecards typically have between 10–15 characteristics
- Static procedure
- Scorecards are out-of-date, even before they are implemented (Hand 2003).

 Application scoring aims at ranking customers – no calibrated probabilities of default needed!

12

Behavioral Scoring

Behavioral Scoring

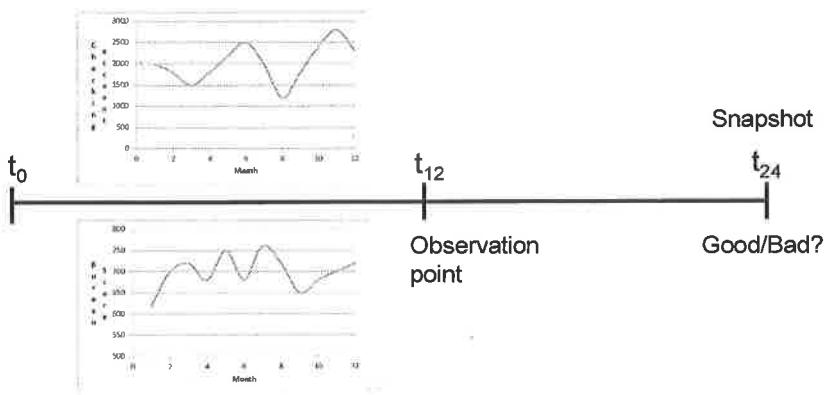
- Study risk behavior of existing customers
- Update risk assessment taking into account recent behavior
- Example of behavior:
 - Average/Max/Min/trend in checking account balance, bureau score, ...
 - Delinquency history (payment arrears, ...)
 - Job changes, home address changes; ...
- Dynamic
- "Video clip" to snapshot (typically taken after 12 months)



13

continued...

Behavioral Scoring



Number of products purchased
Number of times changed home address
Delinquency history (all credits)
...

14

Behavioral Scoring

- Behavioral scoring can be used for:
 - Debt provisioning and profit scoring
 - Authorizing accounts to go in excess
 - Setting credit limits (up- versus down-selling)
 - Renewals/reviews
 - Collection strategies
- Characteristics:
 - Many, many variables (input selection needed, cf. infra)
 - Purpose is again scoring = ranking customers with respect to their default likelihood

16

Dynamic Credit Scoring Models

- Predict default risk for any future point in time (instead of taking snapshot)
- Gives more information than a typical application or behavioral scorecard
- Survival analysis techniques
- Easy to also include changes in economic climate
- Video clip to video clip
- Not being adopted by many financial institutions (yet!)
- First step in profit scoring and/or Customer Lifetime Value (CLV) modeling

16

Corporate Default Risk Modeling

Prediction approach

- Predict bankrupt versus non-bankrupt given ratios (solvency, liquidity, ...) describing financial status of companies
- Assumes the availability of data (and defaults!)
- For example, Altman's z-model for manufacturing firms
 - 1968, linear discriminant analysis
 - For public industrial companies: $Z=1.2X_1 + 1.4X_2 + 3.3X_3 + 0.6X_4 + 1.0X_5$ (healthy if $Z > 2.99$, unhealthy if $Z < 1.81$)
 - For private industrial companies: $Z=6.56X_1 + 3.26X_2 + 6.72X_3 + 1.05X_4$ (healthy if $Z > 2.60$, unhealthy if $Z < 1.1$)
 - X_1 : Working Capital/Total Assets, X_2 : Retained Earnings/Total Assets, X_3 : EBIT/ Total Assets, X_4 : Market (Book) Value of Equity/Total Liabilities, X_5 : Net Sales/Total Assets
 - Extensions to other sectors: z' and z" model

Expert-based approach

- Expert or expert committee decides upon criteria

17

Example: Expert-Based Scorecard

(Ozdemir and Miu 2009)

Business Risk	Score
Industry Position	6
Market Share Trends and Prospects	2
Geographical Diversity of Operations	2
Diversity Product and Services	6
Customer Mix	1
Management Quality and Depth	4
Executive Board Oversight	2
...	...

18

Corporate Default Risk Modeling

Agency rating approach

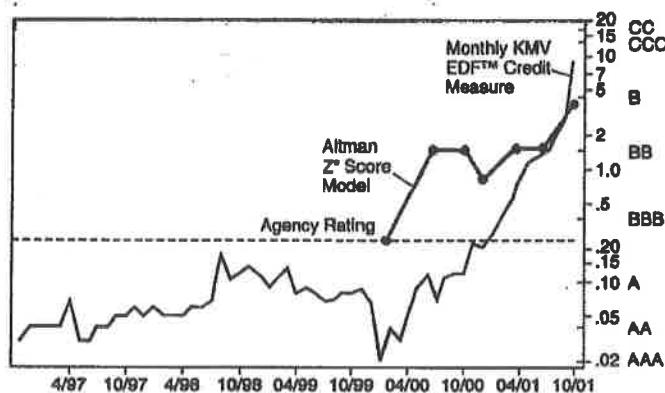
- In the absence of default data (for example, sovereign exposures, low default portfolios)
- Purchase ratings (for example, AAA, AA, A, BBB) to reflect creditworthiness
- Each rating comes with a default probability (PD)

Shadow rating approach

- Purchase ratings from a ratings agency (e.g., Moody's, S&P, Fitch)
- Estimate/mimic ratings using statistical models (e.g., cumulative logistic regression, cf. infra) and data

19

The Altman Model: Enron Case



20

Source: Saunders and Allen (2002)

Rating Agencies

- Moody's Investor Service, Standard & Poor's, Fitch IBCA
- Assign ratings to debt and fixed income securities at the borrower's request to reflect financial strength ("creditworthiness") of the economic entity
- Ratings for
 - Companies (private and public)
 - Countries and governments (sovereign ratings)
 - Local authorities
 - Banks
- Models are based on quantitative and qualitative (judgmental) analysis (black box)

21

Rating Agencies

Moody's	S&P	Fitch	Credit quality
Aaa	AAA	AA+	Extremely strong
Aa1	AA+	AA+	
Aa2	AA	AA	Very strong
Aa3	AA-	AA-	
A1	A+	A+	
A2	A	A	Strong
A3	A-	A-	
Baa1	BBB+	BBB+	
Baa2	BBB	BBB	Adequate
Baa3	BBB-	BBB-	
Ba1	BB+	BB+	
Ba2	BB	BB	Speculative
Ba3	BB-	BB-	
B1	B+	B+	
B2	B	B	Highly speculative
B3	B-	B-	
Caa1	CCC+	CCC+	
Caa2	CCC	CCC	Vulnerable
Caa3	CCC-	CCC-	
Ca	CC	CC	Highly vulnerable
C	C	C	Extremely vulnerable
RD	SD	RD	Selective, restrictive default
D	D	D	Default

22

Van Gestel, Baesens 2008

Chapter 3 Basel I, Basel II, and Basel III

3.1 Basel I, Basel II, and Basel III 3-3

3.1 Basel I, Basel II, and Basel III

Regulatory versus Economic Capital

Role of capital

- Banks should have sufficient capital to protect themselves and their depositors from the risks (credit/market/operational) they are taking.

Regulatory capital

- Amount of capital a bank should have according to a regulation (for example, Basel I, Basel II, or Basel III)

Economic capital

- Amount of capital a bank has based on internal modeling strategy and policy

Actual capital

- Amount of capital a bank actually holds

Types of capital

- **Tier 1:** common stock, preferred stock, retained earnings
- **Tier 2:** revaluation reserves, undisclosed reserves, general provisions, subordinated debt (can vary from country to country)
- **Tier 3:** short-term subordinated debt; specific to market risk (not in Basel III!)

3

The Basel I and II Capital Accords

Basel Committee

- Central Banks/Bank regulators of major (G10) industrial countries
- Meet every three months at Bank for International Settlements (BIS) in Basel

Basel I Capital Accord 1988

- Aim is to set up minimum regulatory capital requirements in order to ensure that banks are able, at all times, to give back depositor's funds.
- Capital ratio (aka Cook ratio) = available capital/risk-weighted assets
- Capital ratio should be > 8%
- Both tier 1 and tier 2 capital

Need for new Accord

- Regulatory arbitrage
- Not sufficient recognition of collateral guarantees
- No differentiation by risk; solvency of debtor not taken into account
- Market risk later introduced in 1996
- No operational risk

4

Basel I: Example

- Minimum capital = 8% of risk-weighted assets
- Fixed risk rates
 - Retail:
 - Cash: 0%
 - Mortgages: 50%
 - Other commercial: 100%
- Example:
 - 100\$ mortgage
 - 50% → 50\$ risk-weighted assets
 - 8% → 4\$ capital needed
- Risk weights are independent of obligor and facility characteristics!

5

Three Pillars of Basel II

- 
- | Pillar 1: Minimum Capital Requirement | Pillar 2: Supervisory Review Process | Pillar 3: Market Discipline and Public Disclosure |
|--|---|--|
| <ul style="list-style-type: none"> ■ Credit Risk <ul style="list-style-type: none"> – Standard Approach – Internal Ratings Based Approach <ul style="list-style-type: none"> ▪ Foundation ▪ Advanced ■ Operational Risk ■ Market Risk | <ul style="list-style-type: none"> ■ Sound internal processes to evaluate risk (ICAAP) ■ Supervisory monitoring | <ul style="list-style-type: none"> ■ Semiannual disclosure of: <ul style="list-style-type: none"> – Bank's risk profile – Qualitative and Quantitative information – Risk management processes – Risk management Strategy ■ Objective is to inform potential investors. |

6

The Standardized Approach

- Risk assessments (for example, AAA, AA, BBB) come from external credit assessment institution (ECAI) (for example, Standard & Poor's, Moody's, Fitch)
- Eligibility criteria for ECAI given in Accord (objectivity, independence, transparency, disclosure, ...)
- Risk weights given in Accord to compute Risk Weighted Assets (RWA)
- Risk weights for sovereigns, banks, corporates, ...
- **Minimum Capital = 0.08 x RWA**

7

The Standardized Approach

For retail:

- Risk weight =75% for non-mortgage; 35% for mortgage

For corporates:

- Risk weights from 20% (AAA) to 150% (B)

For sovereigns:

- Risk weights from 0% (AAA) to 100% (B)

90-day overdue loans weighted at 150%

8

continued...

The Standardized Approach

Example:

- Corporate/1 mio dollars/maturity 5 years/unsecured/S&P rating AA
- RW=20%, RWA=0.2 mio, Reg. Cap.=0.016 mio

Credit risk mitigation (collateral)

- guidelines provided in accord (simple versus comprehensive approach)

But:

- Inconsistencies
- Sufficient coverage?
- Need for individual risk profile!
- No LGD, EAD

9

Internal Ratings Based (IRB) Approach

- Credit Risk: key components
 - Probability of default (PD) (decimal)
 - Loss given default (LGD) (decimal)
 - Exposure at default (EAD) (currency)
 - Maturity (M)
- Expected Loss (EL) = PD x LGD x EAD
- For example, PD=1%, LGD=20%, EAD=1000 Euros → EL=2 Euros
- Standardized Approach
- Internal Ratings Based (IRB) Approach
- In the U.S., initially in the notice of proposed rule making, a difference was made between ELGD (expected LGD) and LGD (based on economic downturn). This was later abandoned.

continued...

10

Probable life of collateral

Internal Ratings Based (IRB) Approach

Internal Ratings Based Approach

- Foundation
- Advanced

	PD	LGD	EAD
Foundation approach	Internal estimate	Regulator's estimate	Regulator's estimate
Advanced approach	Internal estimate	Internal estimate	Internal estimate

11

4th component
Maturity
↳ for risk-based

Worldwide Adoption of Basel II

- 95 nations will implement Basel II in some form by 2015.
- The U.S. implemented Basel II on January 1, 2009.
- The European Union implemented the Accord via the EU Capital Requirements Directives; European banks adopted it in 2008 in parallel with existing system and since 2009 use it exclusively.

12

Adoption of Basel II by BRIC Countries

- Brazil
 - Banco Central do Brasil (Brazilian Central Bank)
 - Standardized approach applied uniformly across all financial institutions operating in Brazil
 - Except larger and internationally active institutions; transition period; first foundation IRB approach and afterwards advanced IRB approach
 - Revised timetable: implementation deadline → 2013

13

continued...

Adoption of Basel II by BRIC Countries

- Russia
 - Банк России (Central Bank of Russia)
 - Implementation of Pillar I was postponed in April 2008 (original deadline: 2009), no clear new deadline
 - Initial focus: implementation of standardized approach
 - Consultation procedures IRB approach being conducted
- India
 - (Reserve Bank of India)
 - Implementation of Basel II standardized approach in 2009
 - Common equity (incl. buffer): 3.6%; Tier 1: 6%; Total Capital: 9 % of risk weighted assets
 - IRB (both foundation as advanced): 2014

14

continued...

Adoption of Basel II by BRIC Countries

- China
 - 中国银行业监督管理委员会 (China Banking Regulatory Commission (CBRC))
 - Six top banks in China must apply Basel II in 2012
 - Large internationally active banks are required to adopt IRB by 2013 at the latest
 - Other banks can choose between Basel II implementation or follow a revised version of the existing capital requirements
 - Banks are required to carry out quarterly compliance assessments

15

Basel II in the U.S.

- Not all U.S. banks subject to Basel II
 - Focus is on the largest and internationally most active banks.
- Core banks required to implement the advanced IRB approach for credit risk.
 - Consolidated total assets of > \$250 billion
 - Consolidated total on-balance sheet foreign exposure > \$10 billion
- Other U.S. banks allowed to voluntarily “opt in” to Basel II advanced IRB approach if they meet the qualification requirements
- Agencies will issue rule similar to standardised Basel II approach for non-core banks

16

IRB Approach

- Split exposure into five categories:
 - corporate (five subclasses)
 - sovereign
 - bank
 - retail (residential mortgage, revolving, other retail exposures)
 - equity
- Foundation IRB approach not allowed for retail
- Risk weight functions to derive capital requirements
- Transition period of three years

17

Basel II: Retail Specifics

- Retail exposures can be pooled.
 - Estimate PD/LGD/EAD per rating/pool/segment!
 - Note that pool=segment=grade=rating=class=cluster.
- Default definition
 - Obligor is past due more than 90 days (can be relaxed during transition period).
 - 180 days in U.K.; 180/120 days in U.S.
 - Default can be applied at the level of the credit facility.
 - Some differences between countries (for example, materiality threshold might vary)!
- PD is the greater of the one-year estimated PD or 0.03%.
- Length of underlying historical observation period to estimate loss characteristics must be at least five years (can be relaxed during three-year transition period, minimum two years at the start).

18

Developing a Rating Based System

- Application and behavioral scoring models provide ranking of customers according to risk.
- This was okay in the past (for example, for loan approval), but Basel II requires well-calibrated default probabilities.
- Map the scores or probabilities to a number of distinct borrower ratings/pools/segments.



- Decide on number of ratings and their definition.
 - Typically around 15 ratings
 - Can be defined by mapping onto Moody's, S&P scale, or expert based (see later)!
 - Impact on regulatory capital!

19

Developing a Rating System

For corporate, sovereign, and bank exposures

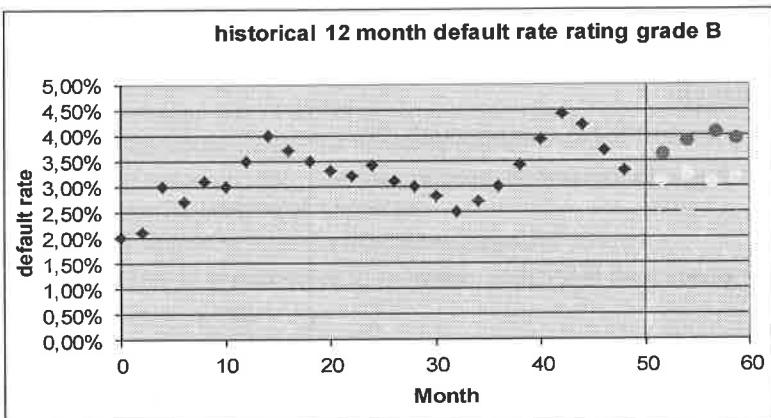
- “*To meet this objective, a bank must have a minimum of seven borrower grades for non-defaulted borrowers and one for those that have defaulted*”, paragraph 404 of the Basel II Capital Accord.

For retail

- “*For each pool identified, the bank must be able to provide quantitative measures of loss characteristics (PD, LGD, and EAD) for that pool. The level of differentiation for IRB purposes must ensure that the number of exposures in a given pool is sufficient so as to allow for meaningful quantification and validation of the loss characteristics at the pool level. There must be a meaningful distribution of borrowers and exposures across pools. A single pool must not include an undue concentration of the bank’s total retail exposure*”, paragraph 409 of the Basel II Capital Accord.

20

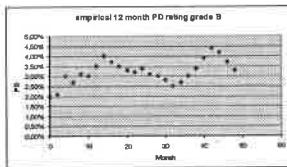
Estimating Well-Calibrated PDs



25

Basel II Model Architecture

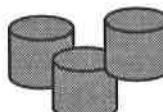
Level 2
Risk ratings definition
PD calibration



Level 1
Model
Scorecard
Logistic Regression

Characteristic Name	Attribute	Scorecard Points
AGE 1	Up to 20	100
AGE 2	21 - 30	120
AGE 3	31 - 37	140
AGE 4	38+	160
GENDER 1	Male	90
GENDER 2	Female	100
SALARY 1	Up to 1000	120
SALARY 2	101 - 1000	140
SALARY 3	1001 - 10000	160
SALARY 4	1001 - 3000	200
SALARY 5	3001+	240

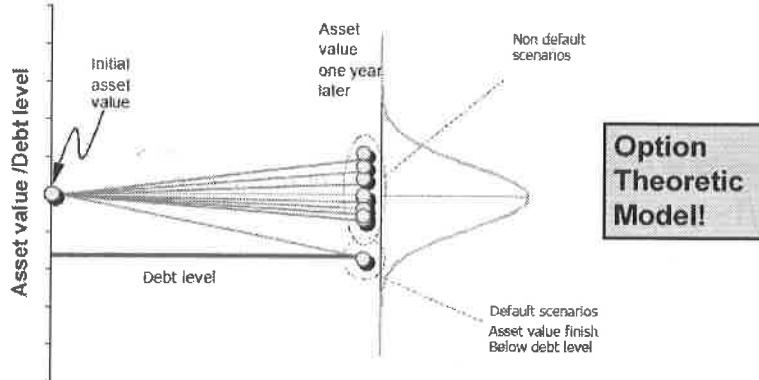
Level 0
Internal Data
External Data
Expert Judgment



26

The Merton Model (1974)

- Obligor defaults when asset value of the firm falls below the debt level at maturity.
- Debt is assumed to be a zero-coupon bond.

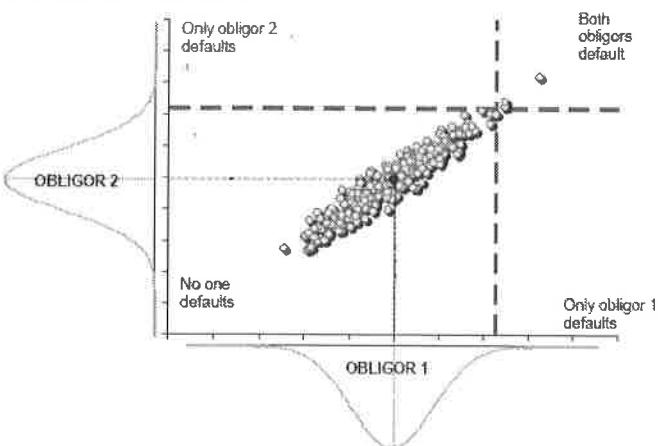


27 Credit Risk Modelling and Basel II, J.C.G. Céspedes, 2002

continued...

The Merton Model

The higher the asset correlation, the higher the probability of common default!



28

The Basel II Model

- Assumptions
 - The number of obligors, n , is very high ($n \Rightarrow \infty$).
 - The exposure size is equal to $1/n$ for all counterparties (goes to 0 for $n \Rightarrow \infty$).
 - All obligors have the same probability of default (model is applied at the level of the rating).
 - Equal pairwise asset correlation ρ between any two obligors, $\text{corr}(A_i, A_j) = \rho$.
- Vasicek assumes asset values for all obligors follow a Gaussian process:

$$A_i = \sqrt{\rho} f + \sqrt{1-\rho} \varepsilon_i$$
- f is a common factor to all companies (one-factor model).
 - For example, economic index (such as world economy)
- ε_i is an idiosyncratic shock.
 - Company-specific risk (for example, related to management quality)

29

continued...

The Basel II Model

- f and ε_i are mutually independent standard normally distributed variables.
- Hence, A_i is also a standard normally distributed variable
- $\text{corr}(A_i, A_j) = \rho$
- The unconditional probability of default becomes: $PD_i = P(A_i \leq D_i) = N(D_i)$
- The conditional probability of default given the systematic factor f becomes:

$$\begin{aligned} PD_{if} &= P(A_i \leq D_i | f) \\ &= P(\sqrt{1-\rho} \varepsilon_i \leq D_i - \sqrt{\rho} f | f) \\ &= P(\varepsilon_i \leq \frac{D_i - \sqrt{\rho} f}{\sqrt{1-\rho}} | f) \\ &= N\left(\frac{D_i - \sqrt{\rho} f}{\sqrt{1-\rho}}\right) \end{aligned}$$
- Given $D_i = N^{-1}(PD_i)$ and put economy to its 99.9% worst state or $N^{-1}(0.001)$, which is equal to $-N^{-1}(0.999)$

$$\alpha^* = PD_{if} = N\left(\frac{N^{-1}(PD_i) + \sqrt{\rho} N^{-1}(0.999)}{\sqrt{1-\rho}}\right)$$

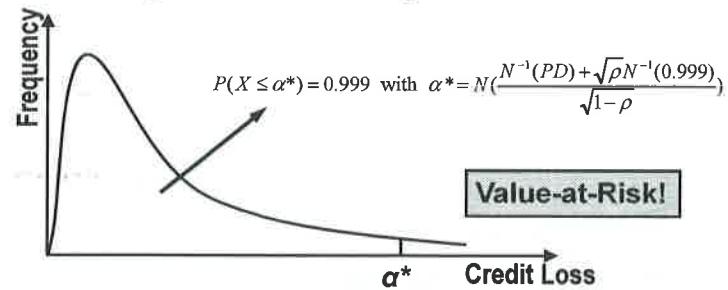
See also Van Gestel, Baesens 2008

30

The Basel II Model

- The corresponding cumulative loss distribution giving the probability that the percentage of defaults X is less than α , $P(X \leq \alpha)$ then becomes (Vasicek 1987, 1991):

$$P(X \leq \alpha) = N \left[\frac{\sqrt{1-\rho} N^{-1}(\alpha) - N^{-1}(PD)}{\sqrt{\rho}} \right]$$

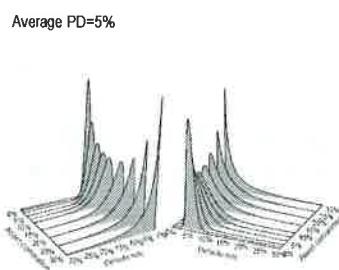
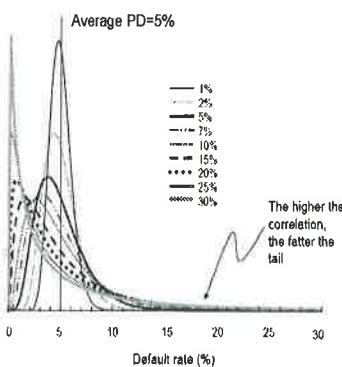


- Can think of α^* as 99.9% worst case PD!

31

continued...

The Basel II Model



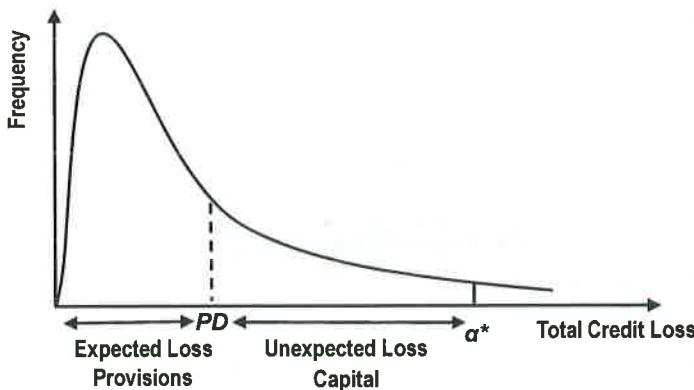
Higher correlation gives fatter tails, and hence higher (unexpected) losses!

32

Credit Risk Modelling and Basel II, J.C.G. Céspedes 2002

The Basel II Value at Risk (VAR) Model

Basel sets α^* at 99.9%, meaning that there is a 0.1% chance (once in 1000 years) that an institution's capital would fail to absorb the unexpected loss and become insolvent!



33

Risk Weight Functions for Retail

Asset condition

$$K = LGD \cdot \left(N\left(\sqrt{\frac{1}{1-\rho}} N^{-1}(PD) + \sqrt{\frac{\rho'}{1-\rho}} N^{-1}(0.999) \right) - PD \right)$$

- $N()$ is cumulative standard normal distribution, $N^{-1}()$ is inverse cumulative standard normal distribution, and ρ is asset correlation.
- Regulatory capital = $K \cdot EAD$
- Residential mortgage exposures $\rho = 0.15$
- Qualifying revolving exposures $\rho = 0.04$
- Other retail exposures

$$\rho = 0.03 \left(\frac{1 - e^{-35PD}}{1 - e^{-35}} \right) + 0.16 \left(1 - \frac{1 - e^{-35PD}}{1 - e^{-35}} \right)$$

34

Risk Weight Functions for Corporates/Sovereigns/Banks

$$K = LGD \left(N \left(\sqrt{\frac{1}{(1-\rho)}} N^{-1}(PD) + \sqrt{\frac{\rho}{(1-\rho)}} N^{-1}(0.999) \right) - PD \right) \frac{(1+(M-2.5)b)}{1-1.5b}$$

$$b = (0.11852 - 0.05478 \ln(PD))^2$$

$$\rho = 0.12 \left(\frac{1-e^{-50PD}}{1-e^{-50}} \right) + 0.24 \left(1 - \frac{1-e^{-50PD}}{1-e^{-50}} \right)$$

- M represents the nominal or effective maturity (between 1 and 5 years)
- Firm size adjustment for SMEs
 - S represents total annual sales in millions of euros (between 5 and 50 million)

$$\rho = 0.12 \left(\frac{1-e^{-50PD}}{1-e^{-50}} \right) + 0.24 \left(1 - \frac{1-e^{-50PD}}{1-e^{-50}} \right) - 0.04 \left(1 - \frac{S-5}{45} \right)$$

35

U.S. Specific Correlations

- Residential mortgages
 - $\rho=0.15$
- Qualifying revolving exposures
 - $\rho=0.04$
- Other retail
 - $\rho = 0.03 + 0.13e^{-35PD}$
- Wholesale (other than HVCRE)
 - $\rho=0.12 + 0.12e^{-50PD}$
- High Volatility Commercial Real Estate (HVCRE)
 - $\rho=0.12 + 0.18e^{-50PD}$

36

Risk Weight Functions

- Regulatory Capital=K (PD, LGD). EAD
- Only covers unexpected loss (UL)!

 - Expected loss=LGD.PD
 - Expected loss covered by provisions!

- RWA not explicitly calculated. Can be backed out as follows:
 - Regulatory Capital=8% . RWA
 - $RWA = 12.50 \times \text{Regulatory Capital} = 12.50 \times K \times EAD$
- Notice the scaling factor of 1.06 for credit-risk-weighted assets (based on QIS 3, 4, and 5).

37

continued...

Risk Weight Functions

- The asset correlations ρ are chosen depending on the business segment (corporate, sovereign, bank, residential mortgages, revolving, other retail) using some empirical but not published procedure!
 - Based on reverse engineering of economic capital models from large banks
 - Asset value correlations reflect a combination of supervisory judgment and empirical evidence (Federal Register)
- ✍ Asset correlations also measure how the asset class is dependent on the state of the economy.

38

LGD/ EAD Errors More Expensive than PD Errors

- Consider a credit card portfolio where
 - $PD = 0.03; LGD = 0.5; EAD = \$10,000$
- Basel formulae give capital requirement of $K(PD, LGD).EAD$

$$K(0.03, 0.50)(10000) = \$343.7$$
- 10% overestimate on PD means capital required is

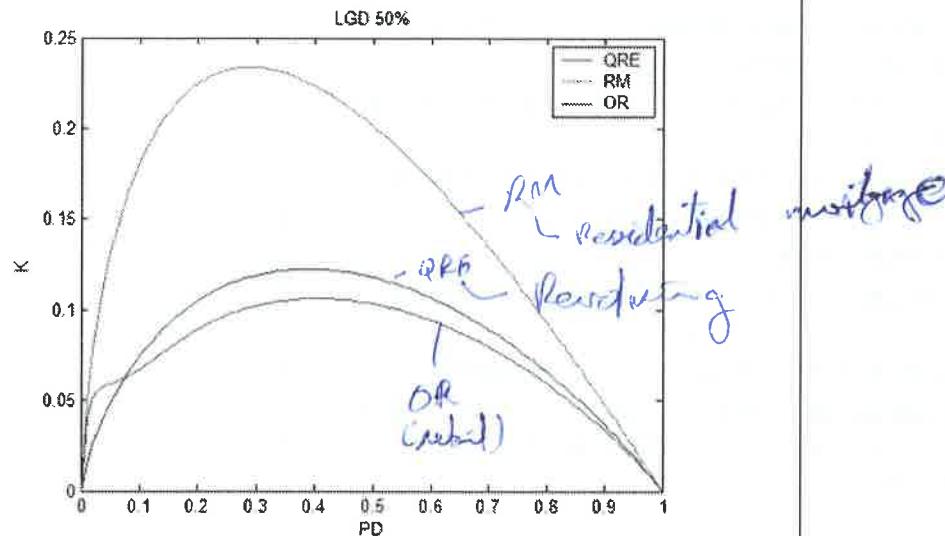
$$K(0.033, 0.50)(10,000) = \$367.3$$
- 10% overestimate on LGD means capital required is

$$K(0.03, 0.55)(10,000) = \$378.0$$
- 10% overestimate on EAD means capital required is

$$K(0.03, 0.50)(11,000) = \$378.0$$

39

The Basel II Risk Weight Functions



40

Basel III

- Response to latest financial crisis
- Objective: Fundamentally strengthen global capital standards
- Key attention points:
 - focus on tangible equity capital because this is the component with the greatest loss-absorbing capacity
 - reduced reliance on banks' internal models
 - reduce the reliance on external ratings
 - greater focus on stress testing

41

continued...

Basel III

- Loss-absorbing capacity beyond common standards for systematically important banks
- Tier 1 must be common shares and retained earnings; Tier 3 eliminated
- Address model risk by, for example, introducing non-risk based leverage ratio as a backstop
- Reduce procyclicality
- Liquidity coverage ratio and the Net Stable Funding Ratio
- No major impact on underlying credit risk models!
- The new standards will take effect on January 1, 2013 and for the most part will become fully effective by January 2019.

42

Basel II versus Basel III

	Basel II	Basel III
Tier 1 capital ratio	4% .RWA	6% .RWA (by 2015)
Core tier 1 capital ratio (common equity=shareholder+ retained earnings)	2% .RWA	4.5% .RWA (by 2015)
Capital conservation buffer (common equity)	-	2,5% .RWA (by 2019)
Countercyclical buffer	-	0% – 2.5% .RWA (by 2019)

Basel III: A global regulatory framework for more resilient banks and banking systems, December 2010, see <http://www.bis.org/publ/bcbs189.pdf>

43

continued...

Basel II versus Basel III

	Basel II	Basel III
Non risk-based leverage ratio (tier 1 capital)	-	3% .Assets (by 2018) Note: Assets include off balance sheet exposures and derivatives.
Total capital ratio	[Tier 1 Capital Ratio] + [Tier 2 Capital Ratio] + [Tier 3 Capital Ratio]	[Tier 1 Capital Ratio] + [Capital Conservation Buffer] + [Countercyclical Capital Buffer] + [Capital for Systemically Important Banks]

Basel III: A global regulatory framework for more resilient banks and banking systems, December 2010, see <http://www.bis.org/publ/bcbs189.pdf>

44

Chapter 4 Preprocessing Data for Credit Scoring and PD Modeling

4.1 Preprocessing Data 4-3

4.1 Preprocessing Data

Motivation

- Dirty, noisy data
 - For example, Age = -2003
- Inconsistent data
 - Value '0' means actual zero or missing value
- Incomplete data
 - Income=?
- Data integration and data merging problems
 - Amounts in euro versus amounts in dollar
- Duplicate data
 - Salary versus professional Income
- Success of the preprocessing steps is crucial to success of following steps
- Garbage in, garbage out (GIGO)
- Very time consuming (80% rule)

3

Preprocessing Data for Credit Scoring

- Types of variables
- Sampling
- Visual data exploration
- Missing values
- Outlier detection and treatment
- Standardizing data
- Transforming data
- Coarse classification and grouping of attributes
- Recoding categorical variables
- Segmentation
- Definition of target variable

4

Types of Variables

- Continuous
 - Defined on a continuous interval
 - For example, income, amount on savings account
 - In SAS: interval
- Discrete
 - Nominal
 - No ordering between values
 - For example, purpose of loan, marital status
 - Ordinal
 - Implicit ordering between values
 - For example, credit rating (AAA is better than AA, AA is better than A, ...)
 - Binary
 - For example, gender

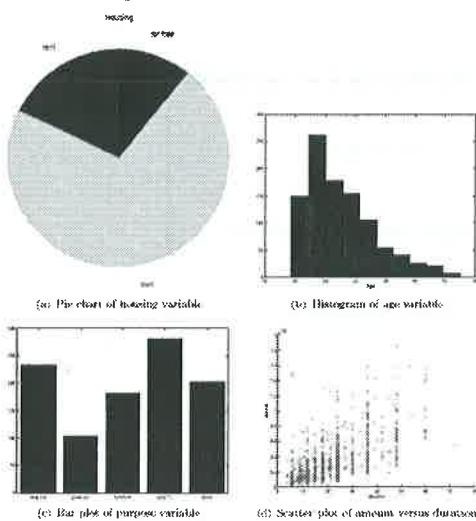
6

Sampling

- Take sample of past applicants to build scoring model
- Think carefully about the population on which the model that is going to be built using the sample will be used
- Timing of sample
 - How far do I go back to get my sample?
 - Trade-off: many data versus recent data
- Number of bads versus number of goods
 - Undersampling, oversampling might be needed (dependent on classification algorithm, see later)
- Sample taken must be from a normal business period to get as accurate a picture as possible of the target population
- Make sure performance window is long enough to stabilize bad rate (for example, 18 months)
- Example sampling problems
 - Application scoring: reject inference
 - Behavioral scoring: seasonality depending upon the choice of the observation point

6

Visual Data Exploration



7

Missing Values

- Reasons
 - Non-applicable (e.g., default date not known for non-defaulters)
 - Not disclosed (e.g., income)
 - Error when merging data (e.g., typos in name and/or ID)
- Keep
 - The fact that a variable is missing can be important information.
 - Add an additional category for the missing values.
 - Add an additional missing value indicator variable (either one per variable, or one for the entire observation).

8

continued...

Missing Values

- Delete
 - When too many missing values, removing the variable or observation might be an option.
 - Horizontally versus vertically missing values
- Replace
 - Estimate missing value using imputation procedures.
 - Be consistent when treating missing values during model development and during model usage!

9

Deleting Missing Values

↓

ID	Age	Income	Marital Status	Credit Bureau Score	Class
1	34	1800	?	620	Bad
2	28	1200	Single	?	Good
3	22	1000	Single	?	Good
4	60	2200	Widowed	700	Bad
5	58	2000	Married	?	Good
6	44	?	?	?	Good
7	22	1200	Single	?	Good
8	26	1500	Married	350	Good
9	34	?	Single	?	Bad
10	50	2100	Divorced	?	Good

10

Imputation Procedures for Missing Values

- For continuous attributes
 - Replace with median/mean (median more robust to outliers)
 - If missing values only occur during model development, can also replace with median/mean of all instances of the same class
- For ordinal/nominal attributes
 - Replace with modal value (= most frequent category)
 - If missing values only occur during model development, replace with modal value of all instances of the same class
- Regression or tree-based imputation
 - Predict missing value using other variables
 - Cannot use target class as predictor if missing values can occur during model usage
 - **More complicated and often do not substantially improve the performance of the scorecard!**

11

Dealing with Missing Values in SAS

```
data credit;
  input income savings;
  datalines;
1300  400
1000 .
2000  800
.      200
1700  600
2500  1000
2200  900
.      1000
1500.
;

proc standard data=credit replace
  out=creditnomissing;
run;
```

Impute node in
SAS Enterprise Miner

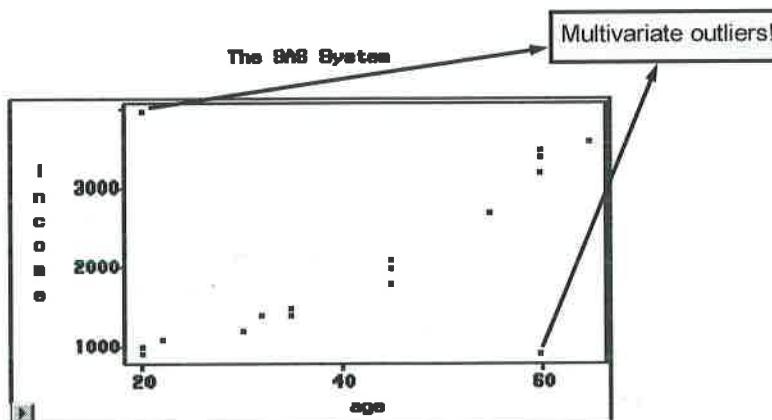
12

Outliers

- Extreme or unusual observations
 - E.g., due to recording, data entry errors or noise
- Types of outliers
 - Valid observation: salary of boss, ratio variables
 - Invalid observation: age = -2003
- Outliers can be hidden in one-dimensional views of the data (multidimensional nature of data)
- Univariate outliers versus multivariate outliers
- Detection versus treatment

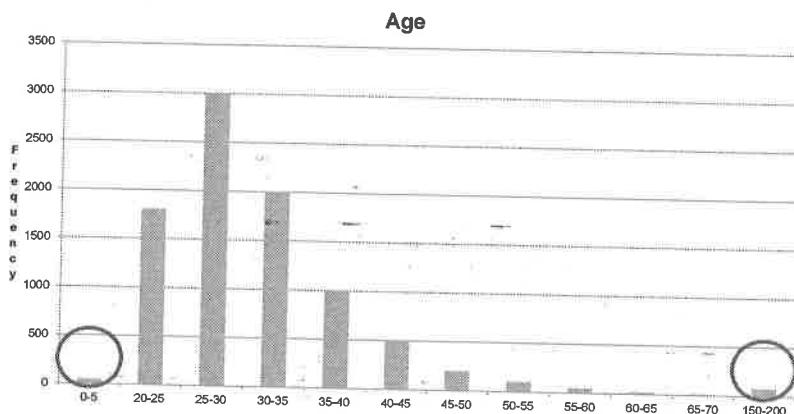
13

Multivariate Outliers



14

Univariate Outlier Detection Methods: Histograms



15

Univariate Outlier Detection Methods: z-score

- z-score measures how many standard deviations an observation lies away from the mean for a specific variable as follows:

$$z_i = \frac{x_i - \mu}{\sigma}$$

- μ is mean of variable x_i and σ standard deviation.
- Outliers are defined when $|z_i| > 3$ (or 2.5).
- The mean of z-scores is 0, standard deviation is 1.
- Calculate the z-score in SAS using PROC STANDARD.

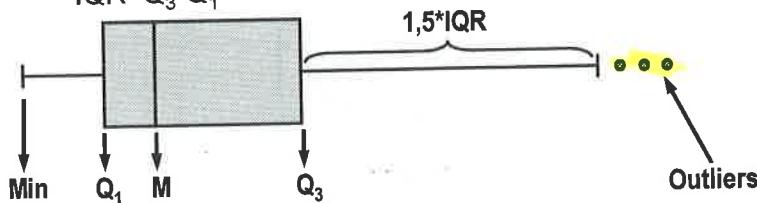
ID	Age	z-score
1	30	(30-40)/10=-1
2	50	(50-40)/10=+1
3	10	(10-40)/10=-3
4	40	(40-40)/10=0
5	60	(60-40)/10=+2
6	80	(80-40)/10=+4

$\mu=40$	$\mu=0$
$\sigma=10$	$\sigma=1$

16

Univariate Outlier Detection Methods: Box Plot

- A box plot is a visual representation of five numbers:
 - Median M $P(X \leq M) = 0.50$
 - First Quartile Q_1 : $P(X \leq Q_1) = 0.25$
 - Third Quartile Q_3 : $P(X \leq Q_3) = 0.75$
 - Minimum
 - Maximum
 - $IQR = Q_3 - Q_1$



17

Multivariate Outlier Detection Methods

- Mahalanobis distance

$$D^2 = (\mathbf{x}_i - \boldsymbol{\mu})^T \boldsymbol{\Sigma}^{-1} (\mathbf{x}_i - \boldsymbol{\mu})$$
 - $\boldsymbol{\mu}$ is the vector of means and $\boldsymbol{\Sigma}$ is the covariance matrix
 - Calculate distance for every point \mathbf{x}_i and sort
- Clustering methods
 - Look for elements outside clusters
- Regression methods
 - Fit regression line and look for points with large errors
 - Residual plots
- Practical advice: only focus on the univariate outliers!

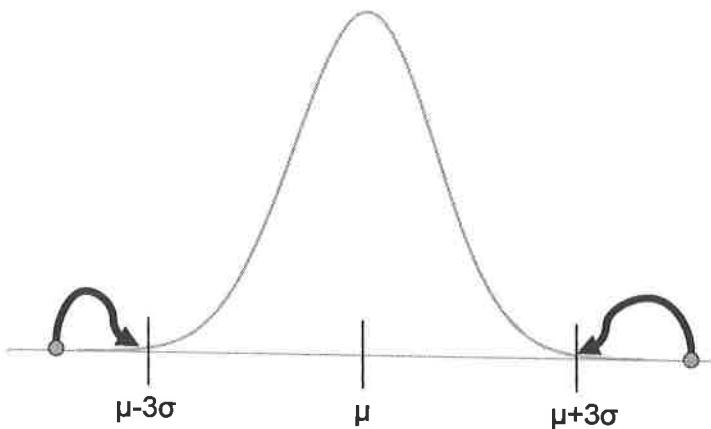
18

Outlier Treatment

- For invalid outliers:
 - For example, age = 300 years
 - Treat as missing value (keep, delete, replace)
- For valid outliers: truncation\winsorizing\capping
 - Truncation based on z-scores:
 - Replace all variable values having z-scores of > 3 with the mean + 3 times the standard deviation
 - Replace all variable values having z-scores of < -3 with the mean -3 times the standard deviation
 - Truncation based on IQR (more robust than z-scores)
 - Truncate to $M \pm 3s$, with $M=\text{median}$ and $s=IQR/(2 \times 0.6745)$
 - See Van Gestel, Baesens et al. 2007
 - Truncation using a sigmoid
 - Use a sigmoid transform, $f(x)=1/(1+e^{-x})$

19

Truncation: Example



20

Standardizing Data

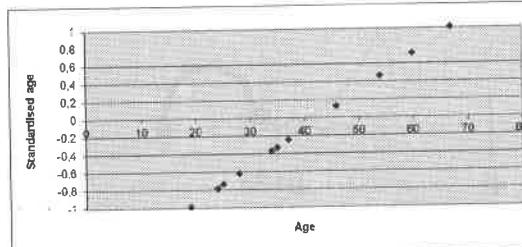
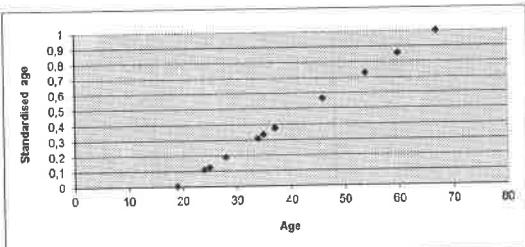
- Motivation
 - Scale variables to the same range (avoid one variable overpowers the other)
 - For example, neural network requires outputs/inputs between 0 and 1
 - For example, salary versus age
- Min/Max standardization

$$X_{new} = \frac{X_{old} - \min(X_{old})}{\max(X_{old}) - \min(X_{old})} (\text{newmax} - \text{newmin}) + \text{newmin}$$
- z-score standardization (when maximum and minimum are outliers)

$$X_{new} = \frac{X_{old} - \text{mean}(X)}{\text{stdev}(X)}$$
- Decimal scaling: $X_{new} = \frac{X_{old}}{10^n}$, with n the number of digits of the maximum absolute value

21

Standardizing Data



22

Computing the z-scores in SAS

```

data credit;
  input age income;
  datalines;
34 1300
24 1000
20 2000
40 2100
54 1700
39 2500
23 2200
34 700
56 1500
;

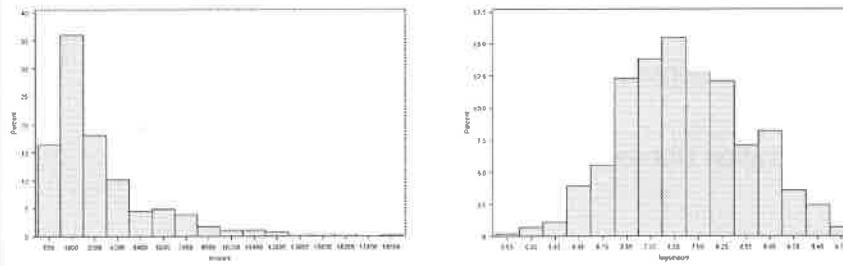
proc standard data=credit mean=0 std=1
  out=creditstand;
run;

```

23

Transforming Data: Log Transform

A logarithmic transform is sometimes adopted to obtain a more symmetric and normal distribution, which is desirable in regression.



Size variables (e.g., assets, loan amount, GDP, ...) are often log transformed as their distribution is typically far from Gaussian.

More complex transforms (e.g., principal component analysis, Box Cox, ...) are also possible but often decrease the interpretability of the variables and scorecard!

24

Coarse Classification and Grouping of Attribute Values

- Motivation
 - Group values of categorical variables for more robust analysis (less dummy variables to be used)
 - Introduce nonlinear effects for continuous variables
 - Classification technique requires discrete inputs (for example, Bayesian network classifiers)
 - Create concept hierarchies: group low-level concepts (for example, raw age data) to higher level concepts (for example, young, middle aged, old)
 - Also called coarse classification, classing, categorization, binning, grouping,...

26

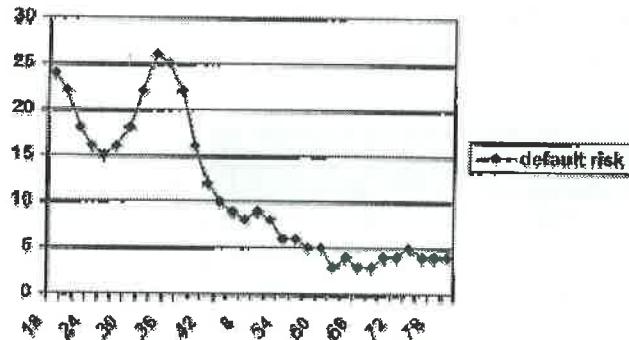
continued...

Coarse Classification and Grouping of Attribute Values

- Methods
 - Equal interval binning
 - Equal frequency binning (histogram equalization)
 - Chi-squared analysis
 - Entropy-based discretization (for example, using decision trees)

26

Coarse Classifying Continuous Variables



Lyn Thomas, "A survey of credit and behavioural scoring: forecasting financial risk of lending to consumers", *International Journal of Forecasting*, 16, 149–172, 2000.

27

The Binning Method

- For example, consider the attribute income
 - 1000, 1200, 1300, 2000, 1800, 1400
- Equal interval binning
 - Bin width=500
 - Bin 1, [1000, 1500[: 1000, 1200, 1300, 1400
 - Bin 2, [1500, 2000[: 1800, 2000
- Equal frequency binning (histogram equalization)
 - 2 bins
 - Bin 1: 1000, 1200, 1300
 - Bin 2: 1400, 1800, 2000
- However, both these methods do not take into account the default risk!

28

The Chi-Squared Method

- Consider the following example (taken from the book *Credit Scoring and Its Applications*, by Thomas, Edelman, and Crook 2002):

Attribute	Owner	Rent Unfurnished	Rent Furnished	With Parents	Other	No Answer	Total
Goods	6000	1600	350	950	90	10	9000
Bads	300	400	140	100	50	10	1000
Good: bad odds	20:1	4:1	2.5:1	9.5:1	1.8:1	1:1	9:1

Suppose we want three categories.

- Should we take option 1 (owners, renters, and others) or option 2 (owners, with parents, and others)?

29

continued...

The Chi-Squared Method

- The number of good owners given that the odds are the same as in the whole population is $(6000+300) \times 9000/10000 = 5670$.
- Likewise, the number of bad renters given that the odds are the same as in the whole population is $(1600+400+350+140) \times 1000/10000 = 249$.
- Compare the observed frequencies with the theoretical frequencies assuming equal odds using a chi-squared test statistic.
- Option 1

$$\chi^2 = \frac{(6000 - 5670)^2}{5670} + \frac{(300 - 630)^2}{630} + \frac{(1950 - 2241)^2}{2241} + \frac{(540 - 249)^2}{249} + \frac{(1050 - 1089)^2}{1089} + \frac{(160 - 121)^2}{121} = 583$$

- Option 2

$$\chi^2 = \frac{(6000 - 5670)^2}{5670} + \frac{(300 - 630)^2}{630} + \frac{(950 - 945)^2}{945} + \frac{(100 - 105)^2}{105} + \frac{(2050 - 2385)^2}{2385} + \frac{(600 - 265)^2}{265} = 662$$

- The higher the test statistic, the better the split (formally, compare with chi-square distribution with $k-1$ degrees of freedom for k classes of the characteristic).
- Option 2 is the better split.

30

Other Methods of Coarse Classification

- Use information gain (from decision tree theory, see later)
 - Find the split/coarse classification that maximizes the information gain
 - Build one-level decision tree (aka decision stump)
- Information Value Statistic (see later)
- Pivot tables

31

Pivot Tables for Coarse Classification



Pivot Table

Customer ID	Age	Purpose	...	G/B		Car	Cash	Travel	Study	House	...
C1	44	car		G		1000	2000	3000	100	5000	
C2	20	cash		G		500	100	200	80	800	
C3	58	travel		B							
C4	26	car		G							
C5	30	study		B							
C6	32	house		G							
C7	48	cash		B							
C8	60	car		G							
...								

Create a pivot table of the attribute versus Good/Bad status and compute the odds.

Group variable values that have similar odds.

32

Coarse Classification in SAS

```
data residence;
  input default$ resstatus$ count;
  datalines;
good owner 6000
good rentunf 1600
good rentfurn 350
good withpar 950
good other 90
good noanswer 10
bad owner 300
bad rentunf 400
bad rentfurn 140
bad withpar 100
bad other 50
bad noanswer 10
;
```

33

Coarse Classification in SAS

```
data coarsel;
  input default$ resstatus$ count;
  datalines;
good owner 6000
good renter 1950
good other 1050
bad owner 300
bad renter 540
bad other 160
;
```

34

Coarse Classification in SAS

```
data coarse2;
  input default$ resstatus$ count;
  datalines;
good owner 6000
good withpar 950
good other 2050
bad owner 300
bad withpar 100
bad other 600
;
```

36

Coarse Classification in SAS

```
proc freq data=coarse1;
  weight count;
  tables default*resstatus / chisq;
run;
```

```
proc freq data=coarse2;
  weight count;
  tables default*resstatus / chisq;
run;
```

36

Recoding Categorical Variables

- Coarse classification reduces the number of dummy indicators needed for the categorical variables.
- Coarse classification introduces new dummy indicators for the continuous variables.
- $Y = \beta_0 + \beta_1 \text{Age}_1 + \beta_2 \text{Age}_2 + \beta_3 \text{Age}_3 + \beta_4 \text{Purp}_1 + \beta_5 \text{Purp}_2 + \beta_6 \text{Purp}_3 + \beta_7 \text{Purp}_4$
- Still need a lot of dummy indicators!
- Look for a monotonic transform $f(\cdot)$ such that
 - $Y = \beta_0 + \beta_1 f(\text{Age}_1, \text{Age}_2, \text{Age}_3) + \beta_2 f(\text{Purp}_1, \text{Purp}_2, \text{Purp}_3, \text{Purp}_4)$.
- Weights of evidence coding

37

Weights of Evidence

- Measures risk represented by each (grouped) attribute category (for example, age 23–26).
- The higher the weight of evidence (in favor of being good), the lower the risk for that category.

Weight of Evidence_{category} = $\ln(p_{\text{good}}_{\text{category}} / p_{\text{bad}}_{\text{category}})$,

where $p_{\text{good}}_{\text{category}} = \text{number of goods}_{\text{category}} / \text{number of goods}_{\text{total}}$
 $p_{\text{bad}}_{\text{category}} = \text{number of bads}_{\text{category}} / \text{number of bads}_{\text{total}}$

If $p_{\text{good}}_{\text{category}} > p_{\text{bad}}_{\text{category}}$ then WOE > 0

If $p_{\text{good}}_{\text{category}} < p_{\text{bad}}_{\text{category}}$ then WOE < 0

38

continued...

Weights of Evidence

Age	Count	Distr. Count	Goods	Distr. Goods	Bads	Distr. Bads	WOE
Missing	50	2.50%	42	2.33%	8	4.12%	-57.28%
18-22	200	10.00%	152	8.42%	48	24.74%	-107.83%
23-26	300	15.00%	246	13.62%	54	27.84%	-71.47%
27-29	450	22.50%	405	22.43%	45	23.20%	-3.38%
30-35	500	25.00%	475	26.30%	25	12.89%	71.34%
35-44	350	17.50%	339	18.77%	11	5.67%	119.71%
44+	150	7.50%	147	8.14%	3	1.55%	166.08%
Total:	2000		1806		194		

Ln [Distr Good / Distr Bad] x 100

39

...

Information Value

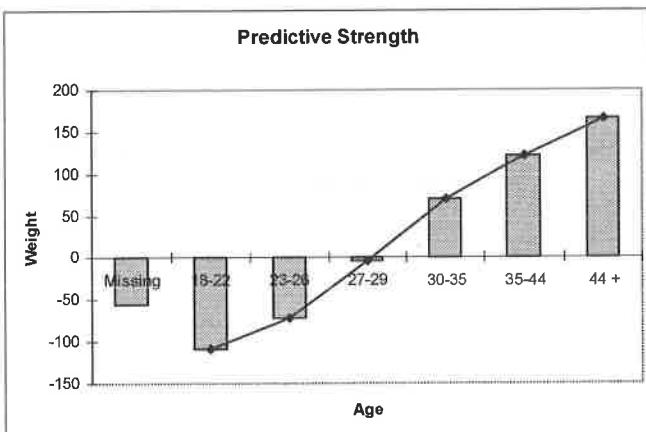
- Information Value (IV) is a measure of predictive power used to
 - assess the appropriateness of the classing
 - select predictive variables.
 - IV is similar to entropy (see later):
- $$\text{IV} = \sum ((p_{\text{good}}_{\text{category}} - p_{\text{bad}}_{\text{category}}) * \text{woe}_{\text{category}})$$
- Rule of thumb:
 - < 0.02 : unpredictable
 - 0.02 – 0.1 : weak
 - 0.1 – 0.3 : medium
 - 0.3 + : strong

Age	Distr. Goods	Distr. Bads	WOE	IV
Missing	2.33%	4.12%	-57.28%	0.0103
18-22	8.42%	24.74%	-107.83%	0.1760
23-26	13.62%	27.84%	-71.47%	0.1016
27-29	22.43%	23.20%	-3.38%	0.0003
30-35	26.30%	12.89%	71.34%	0.0957
35-44	18.77%	5.67%	119.71%	0.1568
44+	8.14%	1.55%	166.08%	0.1095

Information Value: 0.6502

40

WOE: Logical Trend?



Younger people tend to represent a higher risk than the older population.

41

Segmentation

- When more than one scorecard is required
- Build a scorecard for each segment separately
- Three reasons (Thomas, Jo, and Scherer 2001)
 - Strategic
 - Banks might want to adopt special strategies to specific segments of customers (for example, lower cutoff score depending on age or residential status).
 - Operational
 - New customers must have separate scorecard because the characteristics in the standard scorecard do not make sense operationally for them.
 - Variable Interactions
 - If one variable interacts strongly with a number of others, it might be sensible to segment according to this variable.

continued...

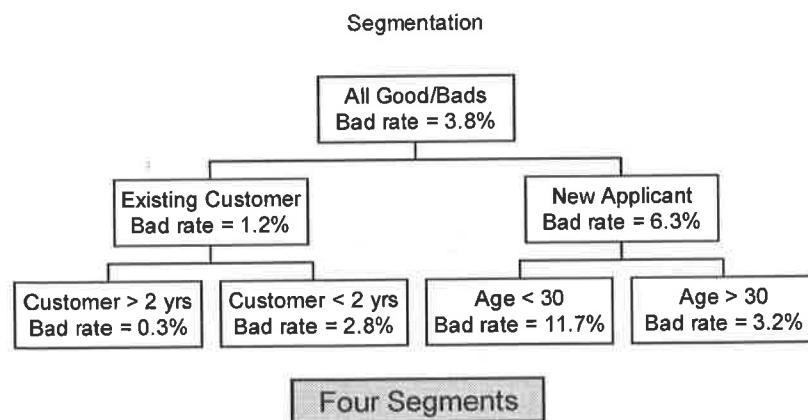
42

Segmentation

- Two ways
 - Experience based
 - In cooperation with financial expert
 - Based on business knowledge and/or experience
 - Statistically based
 - Use clustering algorithms (k-means, decision trees, SOMs, ...)
 - Let the data speak
- Better predictive power than single model if successful!

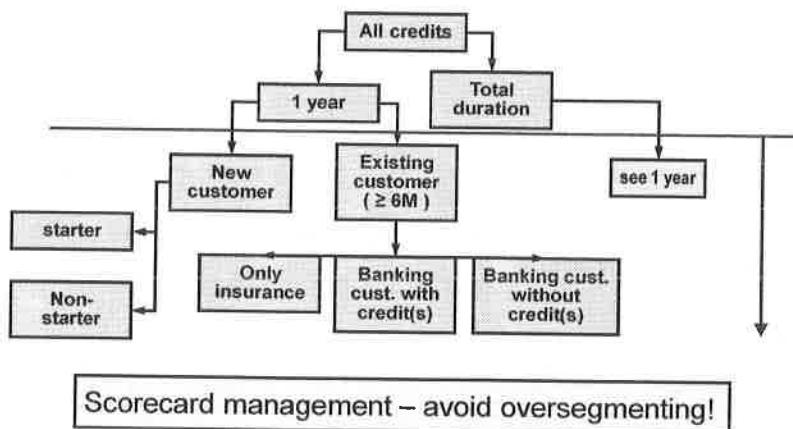
43

Decision Trees for Clustering



44

Segmentation: Example for SMEs



45

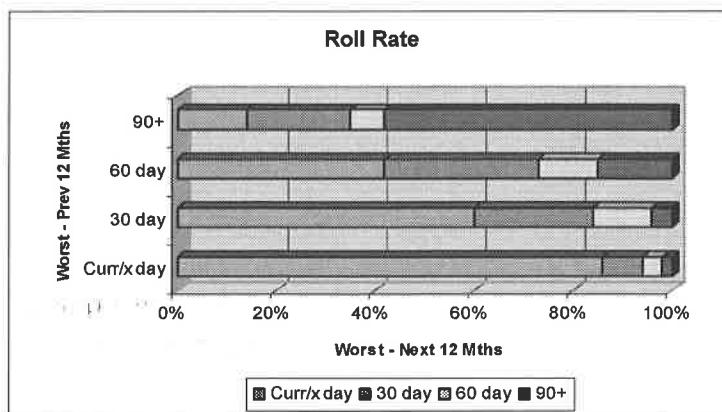
Definitions of Bad

- 30/60/90 days past due
 - Charge off/write off
 - Bankrupt
 - Claim over \$1000
 - Profit based
 - Negative NPV
 - Less than x% owed collected
 - Fraud over \$500
 - Basel II: 90 days, but can be changed by national regulators (for example, U.S., UK, ...)

46

Roll Rate Analysis

Can be used to check stability of bad definition



47

Chapter 5 Classification Techniques for Credit Scoring and PD Modeling

5.1 Classification Techniques 5-3

5.1 Classification Techniques

The Classification Problem

The classification problem can be stated as follows:

Given an observation with characteristics

$x=(x_1, \dots, x_n)$, determine its class c from a predetermined set of classes $\{c_1, \dots, c_m\}$.

- The classes $\{c_1, \dots, c_m\}$ are known beforehand.
- Supervised learning!
- Binary (2 classes) versus multiclass (> 2 classes)

Customer	Age	Income	Gender	...	Good/Bad
John	30	1200	M		Bad
Sarah	25	800	F		Good
Sophie	52	2200	F		Good
David	48	2000	M		Bad
Peter	34	1800	M		Good

3

Methods for Classification

- Statistical methods
- Machine learning methods
- Artificial intelligence methods
- Pattern recognition
- Data mining

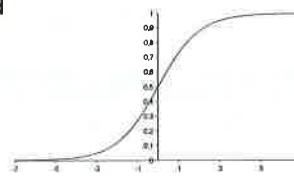
4

Linear Regression for Classification

Customer	Age	Income	Gender	...	Good/Bad	Customer	Age	Income	Gender	...	Good/Bad	Y
John	30	1200	M		Bad	John	30	1200	M		Bad	0
Sarah	25	800	F		Good	Sarah	25	800	F		Good	1
Sophie	52	2200	F		Good	Sophie	52	2200	F		Good	1
David	48	2000	M		Bad	David	48	2000	M		Bad	0
Peter	34	1800	M		Good	Peter	34	1800	M		Good	1

- Linear regression gives: $Y = \beta_0 + \beta_1 \text{Age} + \beta_2 \text{Income} + \beta_3 \text{Gender} + \dots$
- Can be estimated using OLS (in SAS use PROC REG or PROC GLM)
- Two problems:
 - No guarantee that Y is between 0 and 1 (i.e., a probability)
 - Target/Errors not normally distributed
- Use a bounding function to limit the outcome between 0 and 1:

$$f(z) = \frac{1}{1 + e^{-z}}$$



5

Logistic Regression

- Linear regression with a transformation such that the output is always between 0 and 1, and can thus be interpreted as a probability:

$$P(\text{Customer} = \text{good} | \text{age}, \text{income}, \text{gender}, \dots) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 \text{age} + \beta_2 \text{income} + \beta_3 \text{gender} + \dots)}}$$

- The parameters can be estimated in SAS using PROC LOGISTIC.
- After the model has been estimated using historical data, we can use it to score or assign probabilities to new data.

6

Logistic Regression: General Formulation

- The logistic regression model is formulated as follows:

$$P(Y=1|X_1, \dots, X_n) = \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\dots+\beta_nX_n)}} = \frac{e^{\beta_0+\beta_1X_1+\dots+\beta_nX_n}}{1+e^{\beta_0+\beta_1X_1+\dots+\beta_nX_n}}$$

$$P(Y=0|X_1, \dots, X_n) = 1 - P(Y=1|X_1, \dots, X_n) = 1 - \frac{1}{1+e^{-(\beta_0+\beta_1X_1+\dots+\beta_nX_n)}} = \frac{1}{1+e^{\beta_0+\beta_1X_1+\dots+\beta_nX_n}}$$

- Hence,

$$0 \leq P(Y=1|X_1, \dots, X_n), P(Y=0|X_1, \dots, X_n) \leq 1$$

- Model reformulation: $\frac{P(Y=1|X_1, \dots, X_n)}{P(Y=0|X_1, \dots, X_n)} = e^{\beta_0+\beta_1X_1+\dots+\beta_nX_n}$

Ex :-

7

continued...

$$\begin{aligned} P(A) &= 0.75 \\ P(B) &= 0.25 \end{aligned}$$

$$\text{Odds} = \frac{P(A)}{P(B)} = 3:1$$

Logistic Regression

$$\ln\left(\frac{P(Y=1|X_1, \dots, X_n)}{P(Y=0|X_1, \dots, X_n)}\right) = \beta_0 + \beta_1X_1 + \dots + \beta_nX_n$$

$\frac{P(Y=1|X_1, \dots, X_n)}{1-P(Y=1|X_1, \dots, X_n)}$ is the odds in favor of Y=1

$\ln\left(\frac{P(Y=1|X_1, \dots, X_n)}{1-P(Y=1|X_1, \dots, X_n)}\right)$ is called the logit

$$\text{log odds} = \text{logit}$$

8

Interpreting Logistic Regression

■ Odds ratio

- If X_i increases by 1:

$$\text{logit}|_{X_i+1} = \text{logit}|_{X_i} + \beta_i$$

$$\text{odds}|_{X_i+1} = \text{odds}|_{X_i} e^{\beta_i}$$

e^{β_i} is the odds-ratio: the multiplicative increase in the odds when X_i increases by 1 (other variables remaining constant/ceteris paribus)

- $\beta_i > 0 \rightarrow e^{\beta_i} > 1 \rightarrow$ odds/probability increase with X_i
- $\beta_i < 0 \rightarrow e^{\beta_i} < 1 \rightarrow$ odds/probability decrease with X_i

■ Doubling Amount

- The amount of change required for doubling the primary outcome odds
- Doubling amount for X_i equals $\log(2)/\beta_i$

9

Maximum Likelihood Estimation

- Remember: maximum likelihood estimation = maximizing the probability of getting the sample at hand.
- The probability of observing either class is given by

$$P(Y=1|X_1, \dots, X_n)^y (1 - P(Y=1|X_1, \dots, X_n))^{1-y}$$

- Hence, the likelihood of observing the given data set is

$$\prod_{i=1}^N P(Y=1|X_{i1}, \dots, X_{in})^y_i (1 - P(Y=1|X_{i1}, \dots, X_{in}))^{1-y_i}$$

- Take logarithm and optimize using, for example, Newton-Raphson.

10

Hypothesis Testing

- Let L_{full} be the maximum likelihood value of the full model.
- Let $L_{reduced}$ be the maximum likelihood value of the reduced model under the null hypothesis H_0 .
- If L_{full} is almost equal to $L_{reduced}$ then accept H_0 .
- If L_{full} is very different from $L_{reduced}$ then reject H_0 .
- More formally, under H_0

$$-2 \ln \left(\frac{L_{reduced}}{L_{full}} \right) \sim \chi^2_{df}$$

with df the degrees of freedom (number of independent equations in H_0).

- Wald test ($H_0: \beta_i = 0$).

11

Decision Boundary

- Because

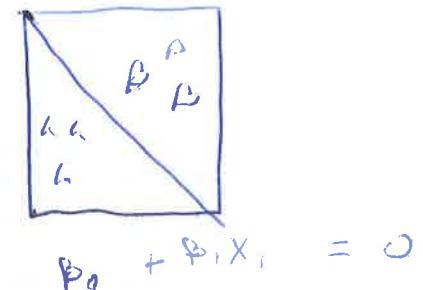
$$\ln \left(\frac{P(Y=1 | X_1, \dots, X_n)}{P(Y=0 | X_1, \dots, X_n)} \right) = \beta_0 + \beta_1 X_1 + \dots + \beta_n X_n$$

the logistic regression classifier assumes a linear decision boundary:

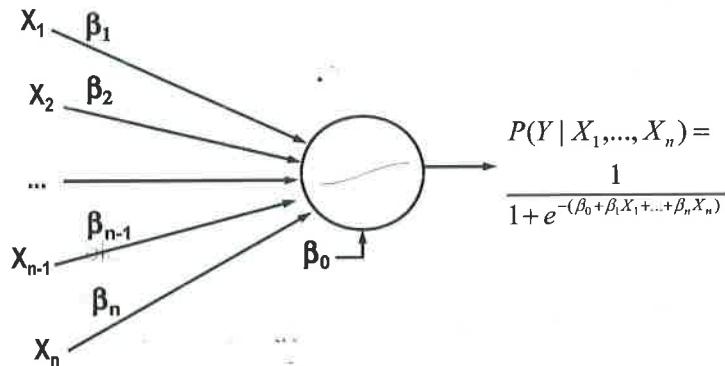
$$\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n = 0$$

- No distributional assumptions (for example, normality) on the independent variables!

12



Neural Network Representation of a Logistic Regression Model



13

Probit and Cloglog Regression

- Probit regression

$$P(Y=1 | X_1, \dots, X_n) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n} e^{-t^2} dt$$

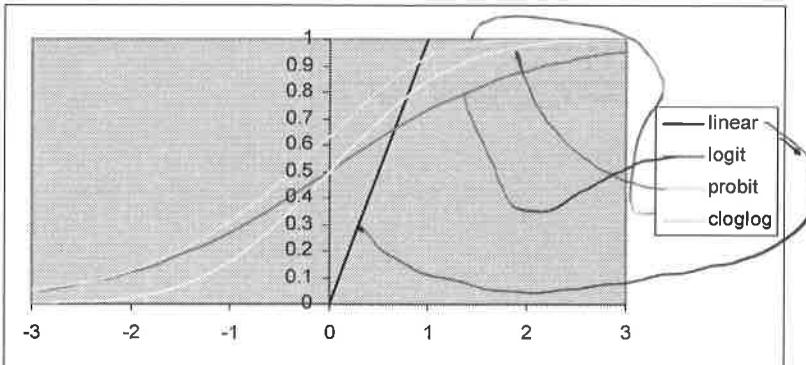
- CDF of standard normal distribution
- Thinner tails than logit model
- For example, Grablowsky and Talley (1981)
- In SAS: PROC LOGISTIC with LINK=PROBIT
- Used in Moody's RiskCalc

- Cloglog regression

- $P(Y=1 | X_1, \dots, X_n) = 1 - \exp(-\exp(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_n X_n))$
- Not symmetric around 0.5
- Increases slowly from 0, but approaches 1 quite suddenly
- In SAS: PROC LOGISTIC with LINK=CLOGLOG

14

Linear/Logistic/Probit/Cloglog Regression



15

Logistic Regression and Weight of Evidence Coding

Actual Age
After classing
After re-coding

Cust ID	Age	Age Group	Age WoE
1	20	1: until 22	-1.1
2	31	2: 22 until 35	0.2
3	49	3: 35+	0.9

16

...

Logistic Regression and Weight of Evidence Coding

$$P(Y=1 | X_1, \dots, X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_{age} * age_woe + \beta_{purpose} * purpose_woe + \dots)}}$$

- No dummy variables!
- More robust

17

Logistic Regression in SAS

Historical data

Customer	Age	Income	Gender	...	Good/Bad
John	30	1200	M		Bad
Sarah	25	800	F		Good
Sophie	52	2200	F		Good
David	48	2000	M		Bad
Peter	34	1800	M		Good



```
proc logistic data=mydata;
  class Gender;
  model Good_Bad=Age Income Gender ...;
run;
```

$$P(\text{Customer} = \text{Good} | \text{Age}, \text{Income}, \text{Gender}, \dots) = \frac{1}{1 + e^{-(0.10 + 0.22\text{age} + 0.650\text{income} - 0.80\text{gender} \dots)}}$$

New data

Customer	Age	Income	Gender	...	Score
Emma	28	1000	F		0,44
Will	44	1500	M		0,76
Dan	30	1200	M		0,18
Bob	58	2400	M		0,88

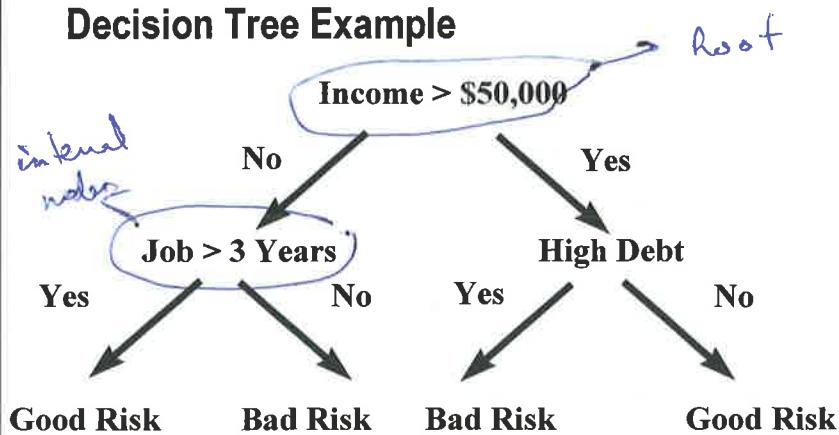
18

Decision Trees

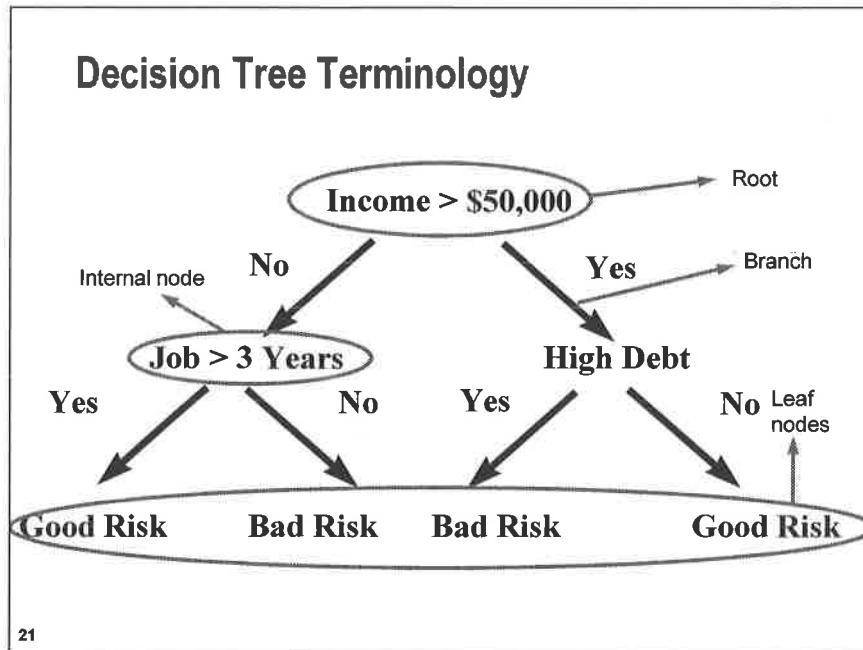
- Recursively partition the training data
- Recursive Partitioning Algorithms (RPAs)
- Various tree induction algorithms
 - C4.5 (See5) (Quinlan 1993)
 - CART: Classification and Regression Trees (Breiman, Friedman, Olshen, and Stone 1984)
 - CHAID: Chi-squared Automatic Interaction Detection (Hartigan 1975)
- Classification tree
 - Target is categorical (e.g., PD)
- Regression tree
 - Target is continuous (interval, e.g., LGD)

19

Decision Tree Example



20



Estimating Decision Trees from Data

- **Splitting decision**
 - Which variable to split at what value (e.g., age < 30 or not, income < 1000 or not; marital status=married or not, ...)?
- **Stopping decision**
 - When to stop growing the tree?
 - When to stop adding nodes to the tree?
- **Assignment decision**
 - Which class (e.g., good or bad customer) to assign to a leave node?
 - Usually the majority class!
 - Can also be based on misclassification costs

22

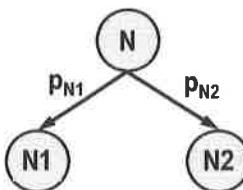
The Splitting Rule

- For each variable, find the best split.
 - For nominal variables, consider splits of the type $X=a, X=b, \dots$
 - For ordinal variables, consider splits of the type $X \leq a$.
 - For continuous (interval) variables, consider splits of the type $X \leq a$.
- Select the best of the best single variable splits.

23

continued...

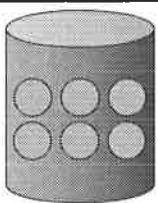
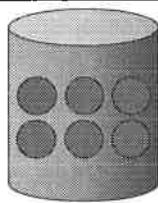
The Splitting Rule



- Choose splits that reduce the impurity of the data.
- A pure node is a node for which all instances are of the same class.
- $I(N)$ is the impurity of node N .
- Goodness of split is weighted mean decrease in impurity $I(N) - p_{N1} I(N_1) - p_{N2} I(N_2)$.
- First approach: choose $I(N)$ as proportion of minority class in node N .
- Other approaches: Entropy and Gini.

24

Splitting Decision: Impurity

Minimal ImpurityMaximal ImpurityMinimal Impurity

- Green dots represent good customers; red dots bad customers.
- Impurity measures disorder/chaos in a data set.
- Impurity of a data sample S can be measured as follows:
 - Entropy: $E(S) = -p_G \log_2(p_G) - p_B \log_2(p_B)$ (C4.5/See5)
 - Gini: $\text{Gini}(S) = 2p_G p_B$ (CART)

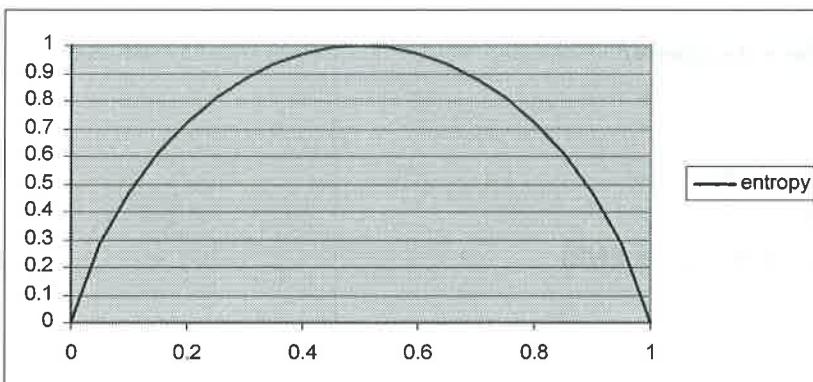
25

The Entropy Measure

- C4.5 (See5)
- Entropy of node N
 - $I(N) = \text{Entropy}(N) = -\sum_j p(j|N) \log_2(p(j|N))$
- Two-class case
 - $\text{Entropy}(N) = -p_1 \log_2(p_1) - p_2 \log_2(p_2)$
 - Notice: $p_1 + p_2 = 1$
- Entropy is maximal when all classes have the same proportion (maximum disorder).
- Entropy is minimal when all observations in N belong to a single class (maximum order).
 - ✍ $-p \ln_2(p)$ approaches 0 when p approaches 0.
- Gain = $\text{Entropy}(N) - p_{N1} \text{Entropy}(N1) - p_{N2} \text{Entropy}(N2)$

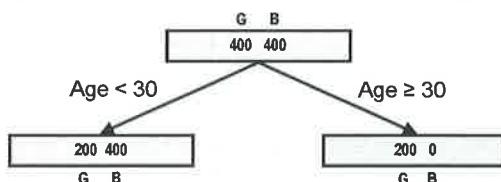
26

The Entropy Measure

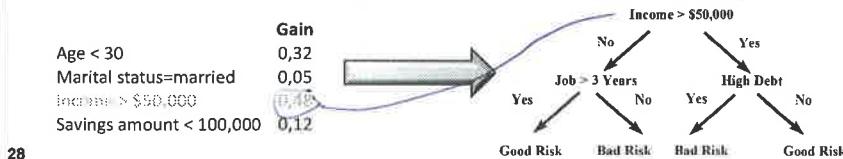


27

Example: Gain Calculation Using Entropy



- Entropy top node = $-1/2 \times \log_2(1/2) - 1/2 \times \log_2(1/2) = 1$
- Entropy left node = $-1/3 \times \log_2(1/3) - 2/3 \times \log_2(2/3) = 0.91$
- Entropy right node = $-1 \times \log_2(1) - 0 \times \log_2(0) = 0$ *impurity*
- Gain = $1 - (600/800) \times 0.91 - (200/800) \times 0 = 0.32$
- Consider different alternative splits and pick the one with biggest gain



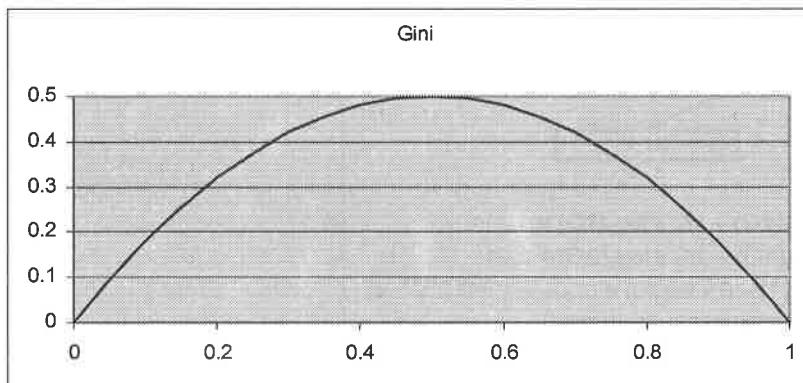
28

The Gini Measure

- CART
 - $I(N) = \sum_{i \neq j} p(i|N)p(j|N) = 1 - \sum_i p(i|N)^2$
- Two-class case
 - $\text{Gini}(S) = 2p_1p_2$
 - Minimal when $p_1=0$ or $p_2=0$
 - Maximal when $p_1=p_2=0.50$
- Gain = $\text{Gini}(N) - p_{N1}\text{Gini}(N1) - p_{N2}\text{Gini}(N2)$

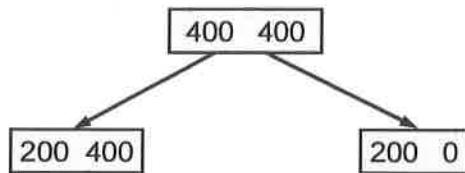
29

The Gini Measure



30

Example: Gain Calculation Using Gini

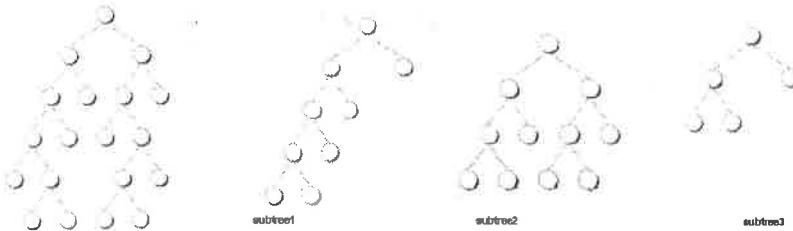


- Gini top node = $2 \times 1/2 \times 1/2 = 0.5$
- Gini left node = $2 \times 1/3 \times 2/3 = 0.44$
- Gini right node = $2 \times 1 \times 0 = 0$
- Gain = $0.5 - (600/800) \times 0.44 - (200/800) \times 0 = 0.166$
- Consider different alternative splits and pick the one with biggest gain
- Typically gives similar splits as when using entropy

31

Stopping Rule

- If you continue splitting, you might end up with one training observation per leaf \Rightarrow the tree is overfitting, too specific for the training data
- Solutions:
 - Don't expand if Gain is below some threshold
 - Grow full tree and post-prune the tree



32

Pruning Decision Trees

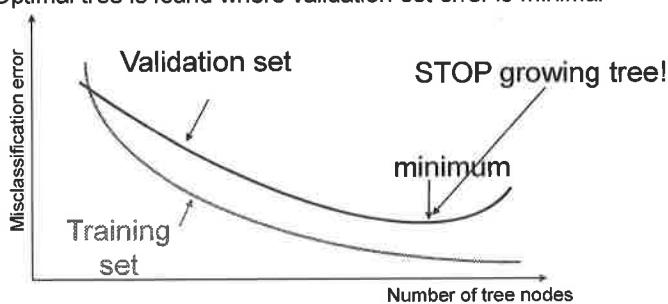
- Split data into a training sample and a validation sample (<> test sample) (typically 70/30)
- Use training sample to grow the tree (to calculate gain)
- Use validation sample to decide on optimal size of the tree
 - Prefer small trees that classify the validation sample best
- Two strategies
 - Grow tree, monitor error on validation set and stop growing when the latter starts to increase
 - Grow full tree, and prune retrospectively using the validation set
- Other methods: for example, C4.5 uses a pruning method based on binomial confidence intervals (retrospectively)

33

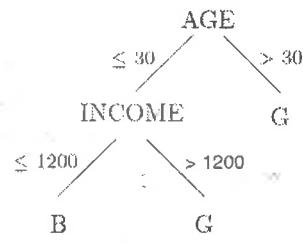
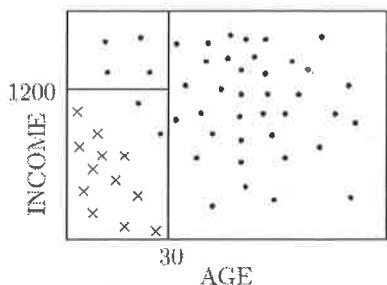
Stopping Decision: Early Stopping

- As the tree is grown, the data becomes more and more segmented and splits are based on fewer and fewer observations
- Need to avoid the tree from becoming too specific and fitting the noise of the data (aka overfitting)
- Prevent overfitting by partitioning the data into a training set (used for splitting decision) and a validation set (used for stopping decision)
- Optimal tree is found where validation set error is minimal

34



Decision Boundary of a Decision Tree



35

Advantages/Disadvantages of Trees

- Advantages
 - Ease of interpretation
 - Nonparametric (no normality assumptions, ...)
 - No need for transformations of variables
 - Robust to outliers
- Disadvantages
 - Sensitive to changes in the training data (unstable)
 - Combinations of decision trees: bagging, boosting, random forests

36

Using Decision Trees for Credit Risk Modeling

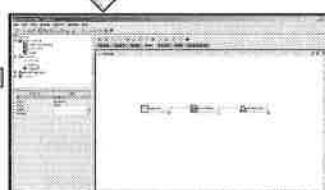
- Alternative to Chi-squared analysis for doing coarse-classification
 - E.g., compute the gain of option 1 (owners, renters, and others) versus option 2 (owners, with parents, and others)
- Screen input variables and do input selection
 - Inputs appearing at the top of the tree are the most important
- Segmentation
 - Build high-level decision tree (e.g., two levels deep) and estimate logistic regression model for each leaf node
- Build decision tree-based scorecard
 - Not advised, as decision trees typically give bad ROC curves (see later)
- Can be used for LGD modeling (regression trees)
 - See later

37

Example Decision Trees in SAS Enterprise Miner

Historical data

Customer	Age	Income	Gender	...	Good/Bad
John	30	1200	M		Bad
Sarah	25	800	F		Good
Sophie	52	2200	F		Good
David	48	2000	M		Bad
Peter	34	1800	M		Good



New data

Customer	Age	Income	Gender	...	Score
Emma	28	1000	F		0,22
Will	44	1500	M		0,60
Dan	30	1200	M		0,10
Bob	58	2400	M		0,92

38

Linear Programming

- Mangasarian (1965)
- Minimize the sum of the absolute values of the deviations (MSD)

$$\begin{aligned} & \min e_1 + e_2 + \dots + e_{n_g} + e_{n_b} \\ & \text{subject to} \\ & w_1x_{i1} + w_2x_{i2} + \dots + w_nx_{in} \geq c - e_i, \quad 1 \leq i \leq n_g, \\ & w_1x_{i1} + w_2x_{i2} + \dots + w_nx_{in} \leq c + e_i, \quad n_g + 1 \leq i \leq n_g + n_b, \\ & e_i \geq 0. \end{aligned}$$

- Use fixed cutoff c , but experiment both with negative and positive c (Freed and Glover 1986)

39

continued...

Linear Programming

- Minimize the maximum deviation (MMD)

$$\begin{aligned} & \min e \\ & \text{subject to} \\ & w_1x_{i1} + w_2x_{i2} + \dots + w_nx_{in} \geq c - e, \quad 1 \leq i \leq n_g, \\ & w_1x_{i1} + w_2x_{i2} + \dots + w_nx_{in} \leq c + e, \quad n_g + 1 \leq i \leq n_g + n_b, \\ & e \geq 0. \end{aligned}$$

- Easy to incorporate a specific bias in an LP model
 - Business experience indicates that variable X_i (e.g., age) is more important than variable X_j (e.g., income)
 - Include constraint $w_i \geq w_j$

40

Integer Programming

$$\min g_1 + g_2 + \dots + g_{n_g} + b_{n_g+1} + b_{n_g+2} + \dots + b_{n_g+n_b}$$

subject to

$$\begin{aligned} w_1 x_{i1} + w_2 x_{i2} + \dots + w_n x_{in} &\geq c - Mg_i, \quad 1 \leq i \leq n_g, \\ w_1 x_{i1} + w_2 x_{i2} + \dots + w_n x_{in} &\leq c + Mb_i, \quad n_g + 1 \leq i \leq n_g + n_b, \\ 0 \leq g_i, b_i &\leq 1 \text{ and integer.} \end{aligned}$$

- M is a constant chosen beforehand, big enough for the problem to be solvable (e.g. 99, 999, 9999, ...)
- g_i and b_i are binary indicators; either 0 or 1
- Value of the objective function then gives the number of misclassifications

41

Linear and Integer Programming

- No statistical underpinning for input selection
- Time consuming
 - Constraint for every data point
 - Integer programming (Branch and Bound methods)
 - Recommended for small samples (typically < 500 Obs)
- Support vector machines!

42

Linear and Integer Programming in SAS

Consider the following data set:

Age	Years at address	Phone	Good/Bad
24	2	No	Bad
35	6	Yes	Bad
40	4	Yes	Bad
30	10	No	Good
28	1	No	Bad
50	5	Yes	Bad
45	15	Yes	Good
60	10	No	Good
20	1	No	Bad
25	5	Yes	Good

43

Linear Programming in SAS (MMD Model)

```
data credit;
input _id_ $ w1 w2 w3 e _type_ $ _rhs_;
datalines;
object 0 0 0 1 min .
const1 24 2 0 -1 LE 100
const2 35 6 1 -1 LE 100
const3 40 4 1 -1 LE 100
const4 30 10 0 1 GE 100
const5 28 1 0 -1 LE 100
const6 50 5 1 -1 LE 100
const7 45 15 1 1 GE 100
const8 60 10 0 1 GE 100
const9 20 1 0 -1 LE 100
const10 25 5 1 1 GE 100
;
proc lp;
run;
```

Scorecard becomes:

If $0.\text{Age} + 9.52 \times \text{Years at Address} + 47.61 \times \text{Phone} > 100$ then
customer =Good

44

w1, w2, e

Integer Programming in SAS

```

data credit2;
input _id_ $ w1 w2 w3 b1 b2 b3 b4 b5 b6 g1 g2 g3 g4 _type_ $ _rhs_ ;
datalines;
object 0 0 1 1 1 1 1 1 1 1 1 min .
const1 24 2 0 -9999 0 0 0 0 0 0 0 0 LE 100
const2 35 6 1 0 -9999 0 0 0 0 0 0 0 LE 100
const3 40 4 1 0 0 -9999 0 0 0 0 0 0 LE 100
const4 30 10 0 0 0 0 0 0 9999 0 0 0 GE 100
const5 28 1 0 0 0 0 -9999 0 0 0 0 0 LE 100
const6 50 5 1 0 0 0 0 -9999 0 0 0 0 LE 100
const7 45 15 1 0 0 0 0 0 0 9999 0 0 GE 100
const8 60 10 0 0 0 0 0 0 0 9999 0 0 LE 100
const9 20 1 0 0 0 0 0 -9999 0 0 0 LE 100
const10 25 5 1 0 0 0 0 0 0 0 9999 GE 100
binary . . 1 1 1 1 1 1 1 1 1 binary
;

proc lp;
run;

```

w2, w3

Variable	Final Value	Start Value	Pivot	Reduced Cost
w2	1	0	1	1
w3	1	0	1	1
b1	0	0	1	0
b2	0	0	1	0
b3	0	0	1	0
b4	0	0	1	0
b5	0	0	1	0
b6	0	0	1	0
g1	0	0	1	0
g2	0	0	1	0
g3	0	0	1	0
g4	0	0	1	0

1 misclassification!

46

Classification Algorithms for Multiclass Data

- Cumulative logistic regression
- Decision trees
- Neural networks
- k-nearest neighbor
- Coding schemes for multiclass data

46

The Cumulative Logistic Regression Model

- When there exists an ordering between the values of the class variable
 - E.g., mapping to external ratings, AAA is better than AA, AA is better than A,
 - $P(C \leq \text{AAA}) \geq P(C \leq \text{AA}) \geq P(C \leq \text{A}) \geq P(C \leq \text{BBB}) \dots$
- In the cumulative logistic regression model, the cumulative probability of corporate C having a rating lower than R, given \mathbf{x} is given by

$$P(C \leq R | \mathbf{x}) = \frac{1}{1 + e^{-\theta_R + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}}$$

$$\frac{P(C \leq R | \mathbf{x})}{1 - P(C \leq R | \mathbf{x})} = e^{-\theta_R + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}$$

$$\log\left(\frac{P(C \leq R | \mathbf{x})}{1 - P(C \leq R | \mathbf{x})}\right) = -\theta_R + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

- Because $P(C \leq \text{AAA})=1$, $\theta_{\text{AAA}}=+\infty$
- Proportional odds model/Parallel logit functions

47

continued...

The Cumulative Logistic Regression Model

- The individual probabilities $P(C=R|\mathbf{x})$ are then obtained as follows:

$$P(C = D | \mathbf{x}) = P(C \leq D | \mathbf{x})$$

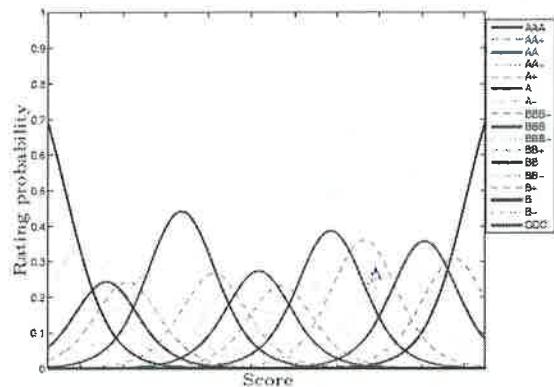
$$P(C = AA | \mathbf{x}) = P(C \leq AA | \mathbf{x}) - P(C \leq D | \mathbf{x})$$

$$P(C = AAA | \mathbf{x}) = 1 - P(C \leq AA | \mathbf{x})$$

- The parameters are estimated using the maximum likelihood principle.
- In SAS: PROC LOGISTIC.

48

Example: Cumulative Logistic Regression

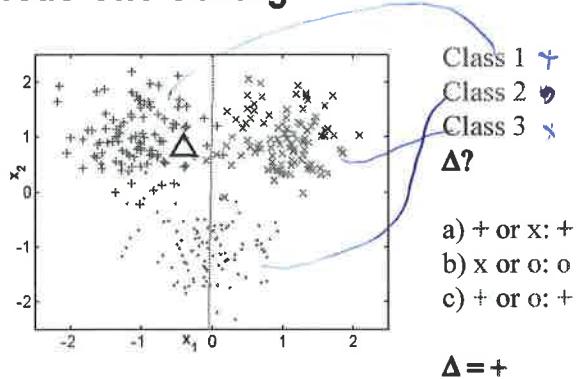


Note: Score = $-\theta_R + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$

Van Gestel T., Martens D., Baesens B., Feremans D., Huysmans J., and Vanthienen J., Forecasting and Analyzing Insurance Companies' Ratings, *International Journal of Forecasting*, 23 (3), pp. 513–529, 2007.

49

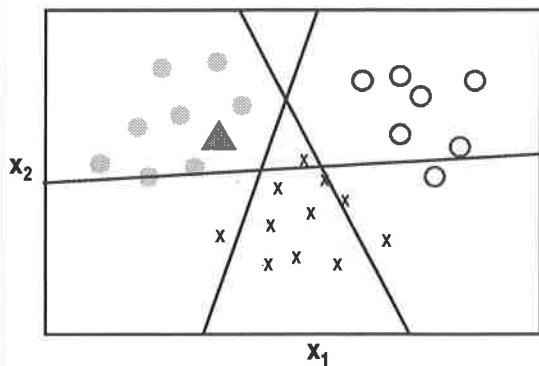
One versus One Coding



T. Van Gestel, *From Linear to Kernel Based Methods for Classification, Modelling and Prediction*, 2002.

50

One versus All Coding



- a) ● or other; $p(\bullet) = 0.92$
- b) ○ or other; $p(\circ) = 0.18$
- c) x or other; $p(x) = 0.30$

Class is ● !

Chapter 6 Measuring the Performance of Credit Scoring Classification Models

6.1 Measuring Performance.....6-3

6.1 Measuring Performance

How to Measure Performance?

- Performance
 - How well does the estimated model perform in predicting new unseen (!) observations?
 - Decide on performance measure
 - Classification: Percentage Correctly Classified (PCC), Sensitivity, Specificity, Area Under ROC curve (AUROC), ...
 - Regression: Mean Absolute Deviation (MAD), Mean Squared Error (MSE), ...
- Methods
 - Split sample method
 - Single sample method
 - N-fold cross-validation

3

Split Sample Method

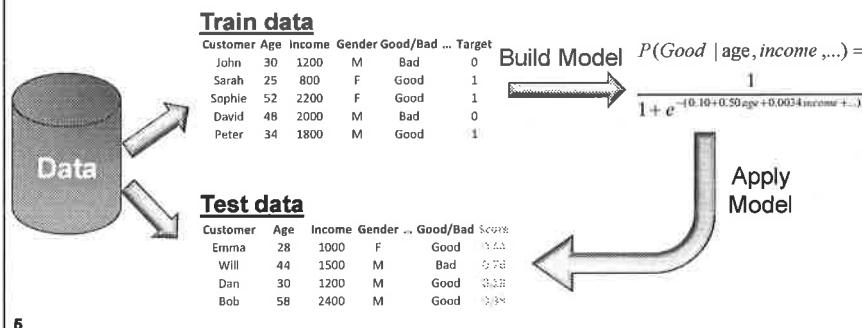
- For large data sets:
 - Large= > 1000 Obs, with > 50 defaults
- Set aside a test set (typically one-third of the data) that is not used during training!
 - ✍ Training set=estimation sample; test set=hold out sample
- Calculate the performance of estimated classifier on the test set.
 - ✍ For decision trees, the validation set is part of the training set.
- Stratification:
 - Same class distribution (good/bad odds) in training set and test set

4

Split Sample Method

Train (Estimation) data versus Test (Hold-out) data

- Train data is used to build model (e.g., logistic regression or decision tree)
- Test data is used to measure performance
- Strict separation between training and test set needed!



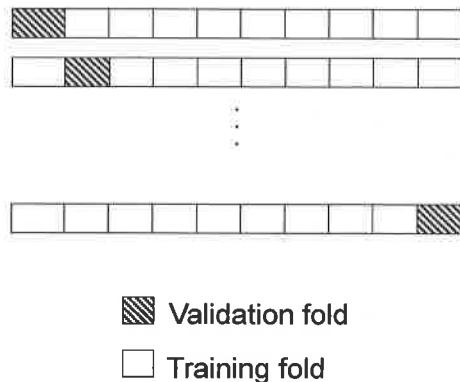
6

N-Fold Cross-Validation

- For small to medium-sized data sets (e.g., < 1000 Obs)
- Split data in N folds (for example, N=10)
- Train on N-1 folds and test on remaining fold
- Repeat N times and compute the mean of the performance measure (can also get standard deviation and/or confidence interval)
- Leave-one-out cross-validation
 - Leave out each observation in turn
 - As many models as observations
- Stratified cross-validation
 - Make sure the good/bad odds are the same in each fold
- Practical advice:
 - Use leave-one-out cross-validation
 - Pick one of the models at random (models differ only in one observation, so very similar anyway), or estimate one model on total data set

6

Example: 10-Fold Cross-Validation



7

Single Sample Method

- For very small data sets
- Performance = $f(\text{training error, model complexity})$
- Penalize for complexity
- For example, Akaike Information Criterion (AIC), Bayesian Information Criterion (BIC)
- Model complexity can be measured by the number of estimated parameters
- Based on statistical learning theory

8

Information Criteria

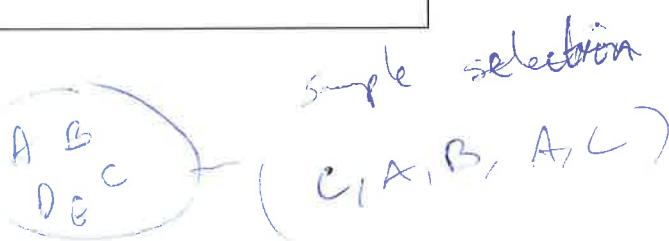
- Akaike Information Criterion (AIC)
 $AIC = -2 \log L + 2 (\text{nr of parameters})$
- Bayesian Information Criterion (BIC)
 $BIC = -2 \log L + (\text{nr of parameters}) * \log(\text{nr of obs})$
 (also known as Schwarz Bayesian Criterion (SBC))
- Best model has minimum AIC or BIC
- Both criteria model the trade-off between the fit of the model and its complexity
- Only meaningful to compare between two models built on the same data set

9

Bootstrapping

- Drawing samples R with repetition from S
- Probability customer is not sampled: $1 - 1/n$
- For n samples: $\left(1 - \frac{1}{n}\right)^n$
 $\lim_{n \rightarrow \infty} \left(1 - \frac{1}{n}\right)^n = e^{-1} = 0.368$
- 0.368 probability that customer does not appear in sample, and 0.632 probability that customer does appear
- Training set = R, Test set = Samples in S but not in R
- Error estimate = 0.368 Error(Training) + 0.632 Error(Test)
- More weight on test set!

10



Performance Measures for Classification

- Confusion matrix, classification accuracy, classification error, sensitivity, specificity
- ROC curve and area under ROC curve
- CAP curve and Accuracy Ratio
- Kolmogorov Smirnov curve and distance
- The Cumulative lift curve
- Multi-class performance measures
 - notch difference graph
 - ROC

11

The Confusion Matrix

		Actual Class	
		Good Payer	Defaulter
Predicted Class	Good payer	True positive (TP)	False positive (FP)
	Defaulter	False Negative (FN)	True Negative (TN)

12

The Confusion Matrix

- Classification accuracy
 $= (TP+TN) / (TP+FP+TN+FN)$
- Error rate $= (FP+FN) / (TP+FP+TN+FN)$
- Sensitivity $= TP / (TP+FN)$
- Specificity $= TN / (TN+FP)$
- All these measures vary when the classification cutoff is varied.
- Extremes
 - Predict all customers as good:
 - Sensitivity=100%, Specificity=0
 - Predict all customers as bad:
 - Sensitivity=0, Specificity=100%

13

Confusion Matrix: Example

	Good/Bad	Score		Good/Bad	Score	Predicted
John	Bad	0,18	Cut off=0,50	John	Bad	0,18
Sophie	Good	0,76	→	Sophie	Good	0,76
David	Bad	0,58		David	Bad	0,58
Emma	Good	0,44		Emma	Good	0,44
Bob	Good	0,84		Bob	Good	0,84

Confusion Matrix

		Actual status	
		Positive (Good)	Negative (Bad)
Predicted status	Positive (Good)	True Positive (Sophie, Bob)	False Positive (David)
	Negative (Bad)	False Negative (Emma)	True Negative (John)

Classification accuracy $= (TP+TN) / (TP+FP+FN+TN) = 3/5$

Classification error $= (FP + FN) / (TP+FP+FN+TN) = 2/5$

Sensitivity $= TP / (TP+FN) = 2/3$

Specificity $= TN / (FP+TN) = 1/2$

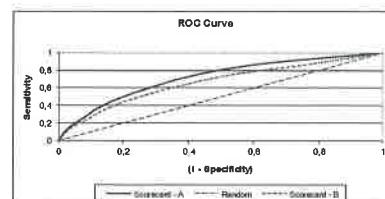
All these measures depend upon the cutoff!

14

The Receiver Operating Characteristic (ROC) Curve

- Make a table with sensitivity and specificity for each possible cutoff
- ROC curve plots sensitivity versus 1-specificity for each possible cutoff

Cut-off	sensitivity	specificity	1-specificity
0	1	0	1
0.01			
0.02			
...			
0.99			
1	0	1	0



- In a credit scoring context, the sensitivity is the percentage of goods predicted to be good, and 1-specificity is the percentage of bads predicted to be good.
- Perfect model has sensitivity of 1 and specificity of 1 (i.e., upper left corner)
- Scorecard A is better than B in above figure

16

good mode

$$0.5 < \text{AUC} < 1$$

bad classification

$$\text{area } \left(\frac{1 \times 1}{2}\right) = 0.5$$

good classification

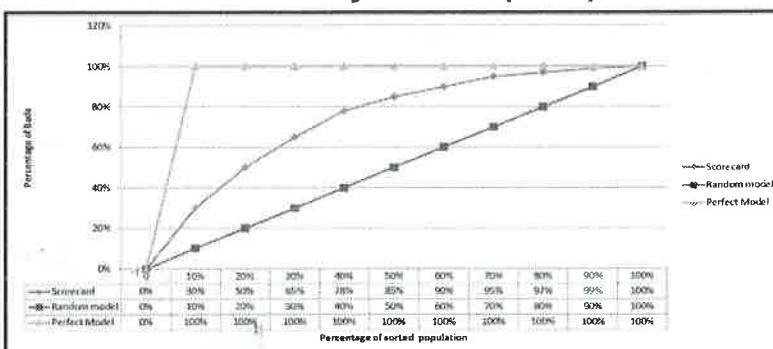
$$\left(\frac{1 \times 1}{1}\right) = 1$$

The ROC Curve

- How to compare intersecting ROC curves?
- ROC curve can be summarized by the area underneath (AUC); the bigger the better!
- The AUC provides a simple figure-of-merit for the performance of the constructed classifier.
- An intuitive interpretation of the AUC is that it provides an estimate of the probability that a randomly chosen instance of class 1 (good payer) is correctly ranked higher than a randomly chosen instance of class 0 (bad payer) (Hanley and McNeil 1983) (Wilcoxon or Mann-Whitney or U statistic).
- A straight line through (0,0) and (1,1) represents a classifier found by randomly guessing the class and serves as a benchmark; hence a good scorecard should have an AUC larger than 0.5.

16

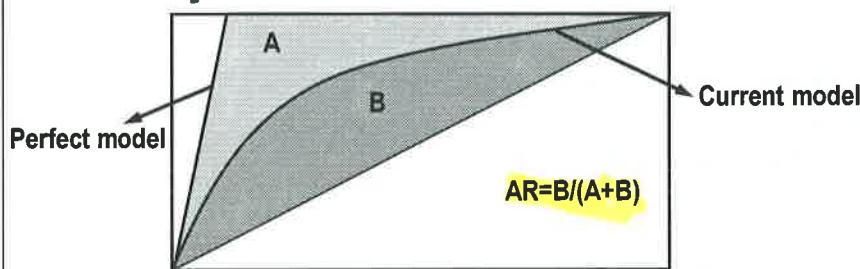
Cumulative Accuracy Profile (CAP)



- Sort the population from high bad score to low bad score.
- Measure the (cumulative) percentage of bads for each score decile.
- E.g., top 30% most likely bads according to model captures 65% of true bads.
- Also referred to as Lorenz curve, Power curve (see Moody's RiskCalc), or Captured Event Plot (SAS).

17

Accuracy Ratio



- The accuracy ratio (AR) is defined as follows:
 $(\text{Area below power curve for current model} - \text{Area below power curve for random model}) / (\text{Area below power curve for perfect model} - \text{Area below power curve for random model})$
- Perfect model has an AR of 1.
- Random model has an AR of 0.
- AR is sometimes also called the Gini coefficient.
- $AR = 2 * AUC - 1$

18

if model is bad

$B = 0$

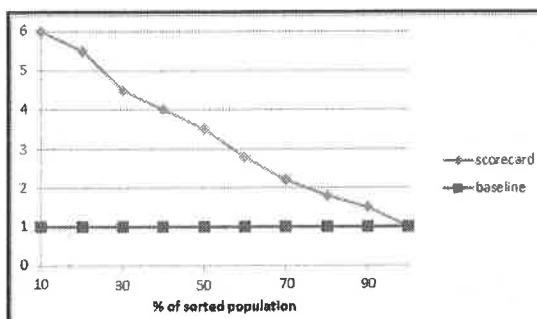
$AR = 0$

if model is good

B tend over 1

$$AR = \frac{1}{D+1} = 1$$

Lift Curve



- Sort population from low score to high score.
- Suppose in top 10% lowest scores, 60% are bads. If in total population 10% are bad, then lift becomes $60/10=6$
- The lift value is thus the cumulative percentage of bads per decile, divided by the overall population percentage of bads.
- Using no model or random sorting, lift would always be 1!
- Lift can also be expressed in a non-cumulative way.

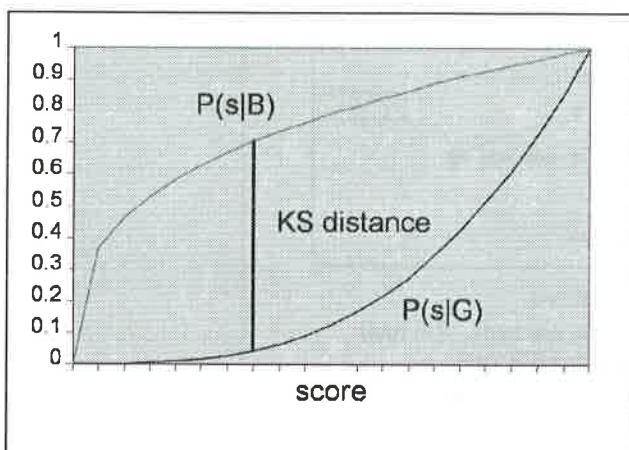
19

The Kolmogorov-Smirnov (KS) Distance

- Separation measure.
- Measures the distance between the cumulative score distributions $P(s|B)$ and $P(s|G)$.
- $KS = \max_s |P(s|G) - P(s|B)|$, where:
 - $P(s|G) = \sum_{x \leq s} p(x|G)$ (equals 1- sensitivity)
 - $P(s|B) = \sum_{x \leq s} p(x|B)$ (equals the specificity)
- KS distance metric is the maximum vertical distance between both curves.
- KS distance can also be measured on the ROC graph:
 - Maximum vertical distance between ROC curve and diagonal

20

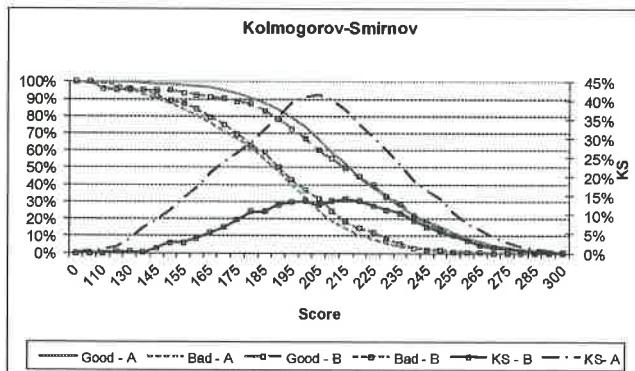
The Kolmogorov-Smirnov Distance



Good model :
bigger KS distance

21

Comparing Scorecards Using the KS Distance



22

...

The Mahalanobis Distance

- Better than Euclidean distance because it takes the distribution (standard deviation) of the scores into account
- Measure the Mahalanobis distance between the two mean scores of the scorecard

$$M = \frac{|\mu_G - \mu_B|}{\sigma}$$

with σ the (pooled) standard deviation of the scores of the goods and the bads from their respective means

- Closely related is the divergence measure D

$$D = \frac{(\mu_G - \mu_B)^2}{\frac{1}{2}(\sigma_G^2 + \sigma_B^2)}$$

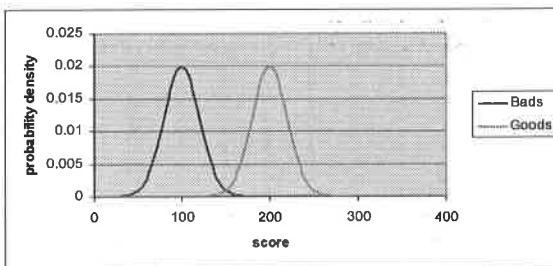
- Only seldomly used!

23

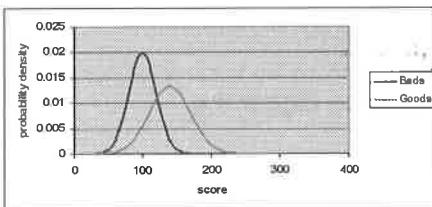
continued...

The Mahalanobis Distance

Good separation



Bad separation



24

Performance Measures for Multi-Class Problems

- For example, rating problems
- Classes are A+, A, A-, B+, B, B-, ...
- Confusion matrix
 - No specificity, sensitivity
- Notch difference graph
- Area under the ROC curve

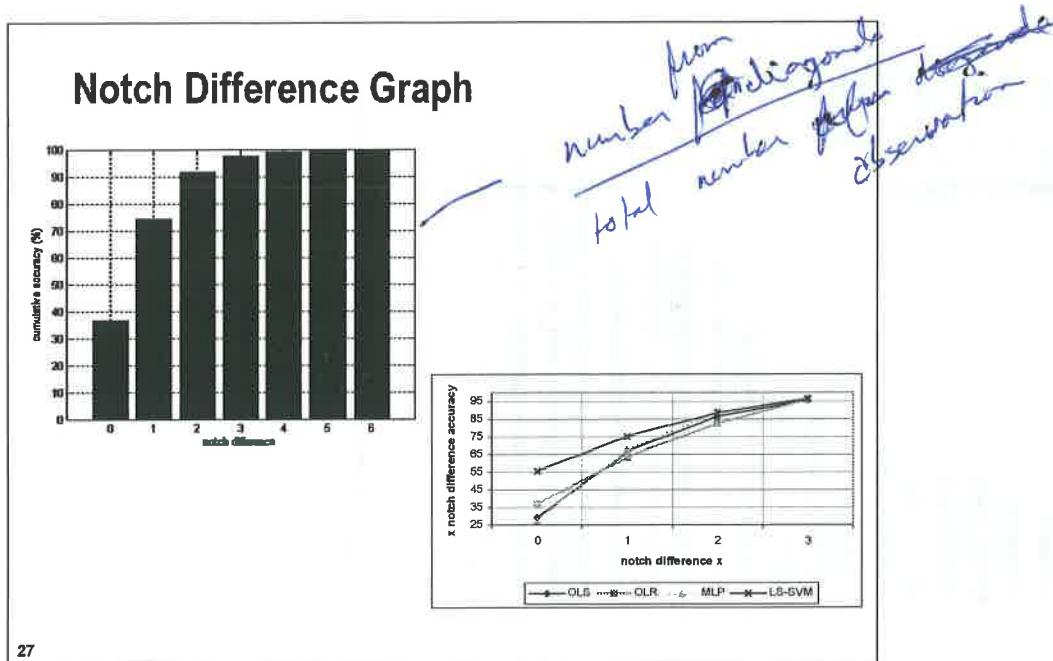
26

Confusion Matrix for Multi-Class Problems

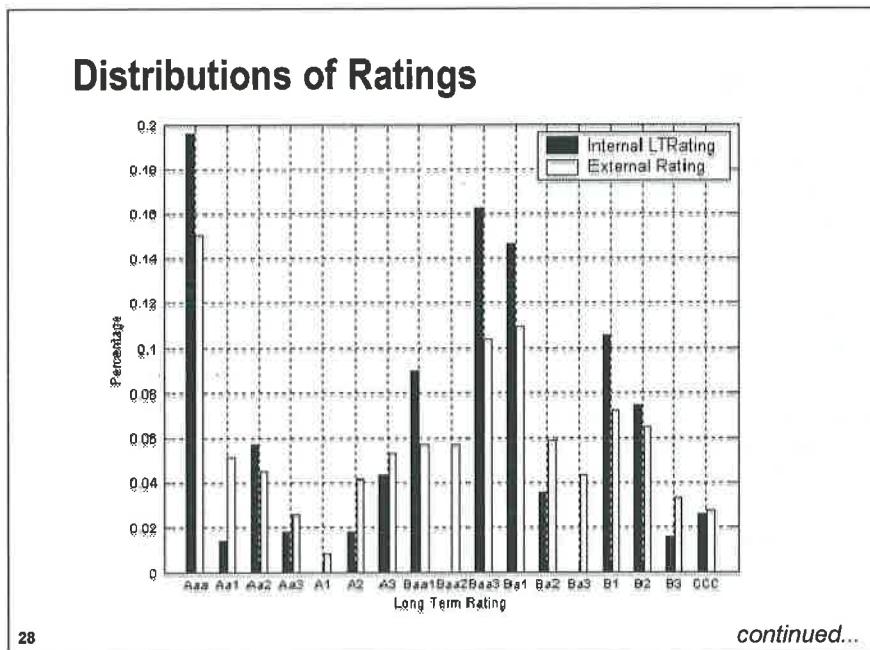
		PREDICTED														
		A+	A	A-	B+	B	B-	C+	C	C-	D+	D	D-	E+	E	E-
TRUE	A+	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
	A	0	14	0	4	0	0	0	0	0	0	0	0	0	0	0
	A-	0	0	0	2	0	0	0	0	0	0	0	0	0	0	0
	B+	0	7	0	60	0	0	5	5	0	1	0	0	0	0	0
	B	0	2	2	54	0	12	58	17	0	5	1	0	1	0	0
	B-	0	0	0	6	0	6	10	6	0	1	0	0	0	0	0
	C+	0	0	0	13	0	3	89	34	0	11	1	0	0	0	0
	C	0	1	0	5	0	1	37	151	1	12	4	0	0	0	0
	C-	0	0	0	2	0	0	3	8	3	4	0	0	0	0	0
	D+	0	0	0	1	0	0	7	20	2	191	24	0	3	1	0
	D	0	0	0	1	0	0	2	9	1	32	122	0	13	7	0
	D-	0	0	0	0	0	0	0	0	0	0	0	2	1	0	0
	E+	0	0	0	0	0	0	0	4	0	12	26	0	37	5	0
	E	0	0	0	0	0	0	0	4	0	4	9	0	5	46	0
	E-	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

- Not all errors have equal impact.
- The further away from the diagonal, the bigger the impact of the error!

26



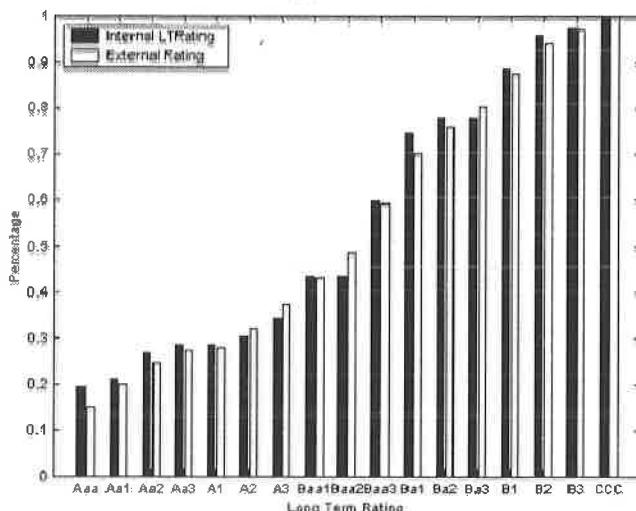
27



28

continued...

Distributions of Rating



29

The AUC for Multiple Classes

- Method 1:
 - Produce n different ROC graphs, one for each class (other classes merged into one) and calculate AUC
 - Calculate overall AUC as weighted average of individual AUCs (use prior probability of class as weights)
- Method 2:
 - Calculate $\binom{n}{2}$ AUCs, one for each possible class comparison and take the average
- References
 - Fawcett, T., *ROC Graphs: Notes and Practical Considerations for Researchers*, HP Labs Tech Report HPL-2003-4.
 - Hand, D. and Till, R.J., "A simple generalization of the area under the ROC curve to multiple class classification problems," *Machine Learning*, 45(2), pp. 171–186, 2001.

30

Chapter 7 Setting the Classification Cutoff

7.1 Setting the Classification Cutoff 7-3

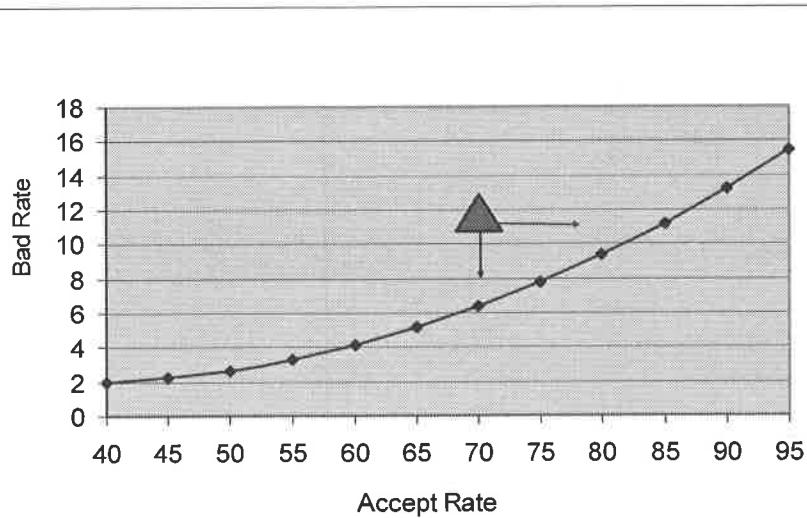
7.1 Setting the Classification Cutoff

Setting the Classification Cutoff

- $P(\text{customer is good payer}|\mathbf{x})$
- Needed in application scoring to make final accept/reject decision
- Depends upon risk preference strategy
- No input provided by Basel accord, capital should be in line with risks taken
- Strategy curve
 - Choose cutoff that produces same acceptance rate as previous scorecard but improves the bad rate
 - Choose cutoff that produces same bad rate as previous scorecard but improves the acceptance rate

3

Strategy Curve



4

Setting the Cutoff Based on Marginal Good-Bad Rates

- Usually one chooses a marginal good-bad rate of about 5:1 or 3:1
- Dependent upon the granularity of the cutoff change

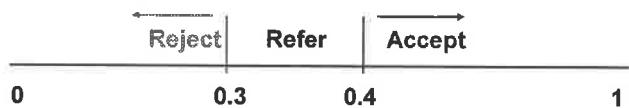
Cut-off	Cum. Goods above score	Cum. Bads above score	Marginal good-bad rate
0.9	1400	50	-
0.8	2000	100	12:1
0.7	2700	170	10:1
0.6	2950	220	5:1
0.5	3130	280	3:1
0.4	3170	300	2:1

6

Including a Refer Option

Refer option

- Gray zone
- Requires further (human) inspection



6

...



Case Study 1

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., and Vanthienen, J., "Benchmarking State of the Art Classification Algorithms for Credit Scoring," *Journal of the Operational Research Society*, Volume 54, Number 6, pp. 627–635, 2003.

7

Benchmarking Study Baesens et al. 2003

- Eight real-life application credit scoring data sets
 - U.K., Benelux, Internet
- Seventeen algorithms or variations of algorithms
- Various cutoff setting schemes
- Classification accuracy + Area under Receiver Operating Characteristic Curve
- McNemar test + DeLong, DeLong and Clarke-Pearson test

Baesens, B., Van Gestel, T., Viaene, S., Stepanova, M., Suykens, J., Vanthienen, J., "Benchmarking State of the Art Classification Algorithms for Credit Scoring," *Journal of the Operational Research Society*, Volume 54, Number 6, pp. 627–635, 2003.

8

continued...

Benchmarking Study

Table 1 Characteristics of credit scoring data sets

	<i>Inputs</i>	<i>Data set size</i>	<i>Training set size</i>	<i>Test set size</i>	<i>Goodbank</i>
Bend	33	3123	2082	1041	66.7/33.3
Bone2	33	2190	1493	2397	70/30
UK1	16	9360	6240	3120	75/25
UK2	16	11709	7800	3909	89/10
UK3	19	3960	2640	1320	90/10
UK4	19	1980	1320	660	80/20
Geru	20	1000	666	334	70/30
Aust	14	690	460	230	55.5/44.5

9

Benchmarking Study: Conclusions

- Neural networks and support vector machines consistently yield good performance in terms of both classification accuracy and area under ROC curve
- Simple linear classifiers such as logistic regression also gave very good performances, most often not significantly different from neural network performance
- Most credit scoring data sets are only weakly nonlinear
- Flat maximum effect

10

continued...

Benchmarking Study: Conclusions

- “Although the differences are small, they may be large enough to have commercial implications.” (Henley and Hand 1997)
- “Credit is an industry where improvements, even when small, can represent vast profits if such improvement can be sustained.” (Kelly 1998)
- For mortgage portfolios, a small increase in PD scorecard discrimination will significantly lower the minimum Capital Requirements in the context of Basel II.
- The best way to augment the performance of a scorecard is to improve data quality (e.g., by looking for better predictors).
- Role of credit bureaus!

11

Credit Bureaus

- Credit reference agencies or credit bureaus
- U.K.: information typically accessed through name and address
- Types of information:
 - Publicly available information
 - For example, time at address, electoral rolls, court judgments
 - Previous searches from other financial institutions
 - Shared contributed information by several financial institutions
 - Check whether applicant has loan elsewhere and how he pays
 - Aggregated information
 - For example, data at ZIP code level
 - Fraud Warnings
 - Check whether prior fraud warnings at given address
 - Bureau added value
 - For example, build generic scorecard for small populations or new product
- In U.K.: Experian, Equifax, Call Credit
- In U.S.: Equifax, Experian, TransUnion

12

Evaluating Scorecards

- **Statistical performance**
 - Model discrimination: AUC, Gini
 - Model calibration: binomial, chi-squared tests, ... (see later)
- **Interpretability + Justifiability**
 - Very subjective, but *crucial!*
 - Often need to be balanced against statistical performance
- **Operational efficiency**
 - How much effort is needed to evaluate/monitor/retrain the scorecard?
- **Economical cost**
 - What is the cost to gather the model inputs and evaluate the scorecard?
 - Is it worthwhile buying external data and/or models (e.g., FICO)?
- **Regulatory compliance**
 - In accordance with regulation and legislation
 - E.g., Basel II, Solvency II

Chapter 8 Input Selection for Classification

8.1 Input Selection..... 8-3

8.1 Input Selection

Input Selection

- Inputs=Features=Attributes=Characteristics=Variables.
- Also called feature selection, attribute selection, characteristic selection, variable selection
- If n features are present, $2^n - 1$ possible feature sets can be considered.
- Heuristic search methods are needed!
- Good feature subsets contain features highly correlated with (predictive of) the class, yet uncorrelated with (not predictive of) each other (Hall and Smith 1998).
- Can improve the performance and estimation time of the classifier.

3

continued...

Input Selection

- Curse of dimensionality
 - The number of training observations required increases exponentially with the number of inputs.
 - Remove irrelevant and redundant inputs.
- Interaction and correlation effects
- Correlation implies redundancy
- Difference between redundant input and relevant input
 - An input can be redundant, but that does not mean it is not relevant (due to, for example, high correlation with another input already in the model).
- Example: correlation between Age and Time on Books

4

Input Selection Procedure

- Step 1: Use a filter procedure
 - For quick filtering of inputs
 - Inputs are selected independent of the classification algorithm (e.g., logistic regression).
- Step 2: Forward/backward/stepwise regression
 - Use the p -value of the logistic regression for input selection.
- Step 3: AUC-based pruning
 - Iterative procedure based on AUC

* information value to filter *

5

Filter Methods for Input Selection

	Continuous target (e.g., LGD)	Discrete target (e.g., PD)
Continuous input (e.g., income)	Pearson correlation	Fisher score
Categorical input (e.g., marital status)	Fisher score ANOVA analysis	Chi-squared analysis Cramer's V Information value Gain/entropy

6

Pearson Correlation

- Compute Pearson correlation between each continuous variable and continuous target
- Always varies between -1 and +1
- Only keep variables for which $|\rho_P| > 0.50$; or keep, e.g., top 10%

7

Chi-Squared-Based Filter

<u>Observed Frequencies</u>		Good Payer	Bad Payer	Total
	Married	500	100	600
	Not Married	300	100	400
	800	200	1000	

Under the independence assumption,
 $P(\text{married and good payer}) = P(\text{married}).P(\text{good payer}) = 0.6 * 0.8$.
The expected number of good payers that are married is $0.6 * 0.8 * 1000 = 480$.

<u>Independence Frequencies</u>		Good Payer	Bad Payer	Total
	Married	480	120	600
	Not Married	320	80	400
	800	200	1000	

8

continued...

Forward/Backward/Stepwise Regression

- Use the p-value to decide upon the importance of the inputs:
 - $p\text{-value} < 0.01$: highly significant
 - $0.01 < p\text{-value} < 0.05$: significant
 - $0.05 < p\text{-value} < 0.10$: weakly significant
 - $0.1 < p\text{-value}$: not significant
- Can be used in different ways:
 - Forward
 - Backward
 - Stepwise

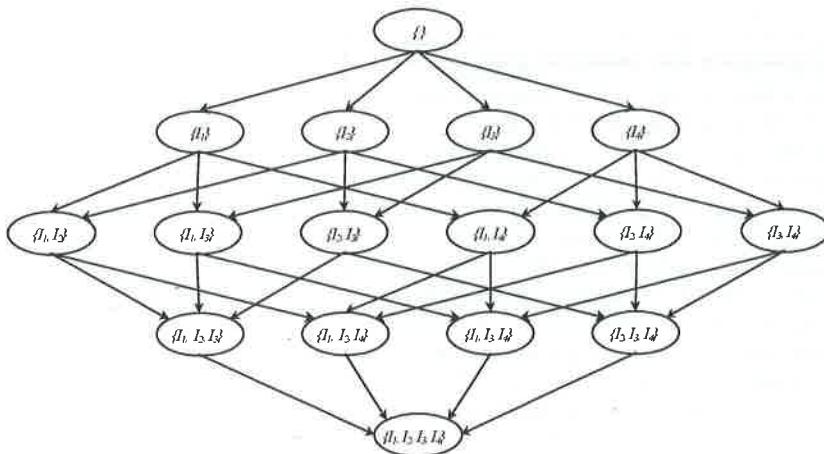
13

Search Strategies

- Forward selection
 - Starts from empty model and always adds variables based on low p-values
- Backward elimination
 - Starts from full model and always deletes variables based on high p-values
- Stepwise
 - Starts as forward selection, but checks whether added variables cannot be removed later

14

Example: Search Space for Four Inputs



Note: $I_1 = \text{age}$; $I_2 = \text{income}$;
 $I_3 = \text{marital status}$; $I_4 = \text{employment status}$

16

Forward/Backward/Stepwise Logistic Regression

SELECTION= FORWARD

PROC LOGISTIC first estimates parameters for effects forced into the model. These effects are the intercepts and the first n explanatory effects in the MODEL statement, where n is the number specified by the START= or INCLUDE= option in the MODEL statement (n is zero by default). Next, the procedure computes the score chi-square statistic for each effect not in the model and examines the largest of these statistics. If it is significant at the SLENTRY= level, the corresponding effect is added to the model. After an effect is entered in the model, it is never removed from the model. The process is repeated until none of the remaining effects meet the specified level for entry or until the STOP= value is reached.

16

continued...

Stepwise Logistic Regression

SELECTION=BACKWARD

Parameters for the complete model as specified in the MODEL statement are estimated unless the START= option is specified. In that case, only the parameters for the intercepts and the first n explanatory effects in the MODEL statement are estimated, where n is the number specified by the START= option. Results of the Wald test for individual parameters are examined. The least significant effect that does not meet the SLSTAY= level for staying in the model is removed. After an effect is removed from the model, it remains excluded. The process is repeated until no other effect in the model meets the specified level for removal or until the STOP= value is reached.

17

continued...

Stepwise Logistic Regression

SELECTION=STEPWISE

This is similar to the SELECTION=FORWARD option except that effects already in the model do not necessarily remain. Effects are entered into and removed from the model in such a way that each forward selection step might be followed by one or more backward elimination steps. The stepwise selection process terminates if no further effect can be added to the model or if the effect just entered into the model is the only effect removed in the subsequent backward elimination.

SELECTION=SCORE

PROC LOGISTIC uses the branch and bound algorithm of Furnival and Wilson (1974).

18

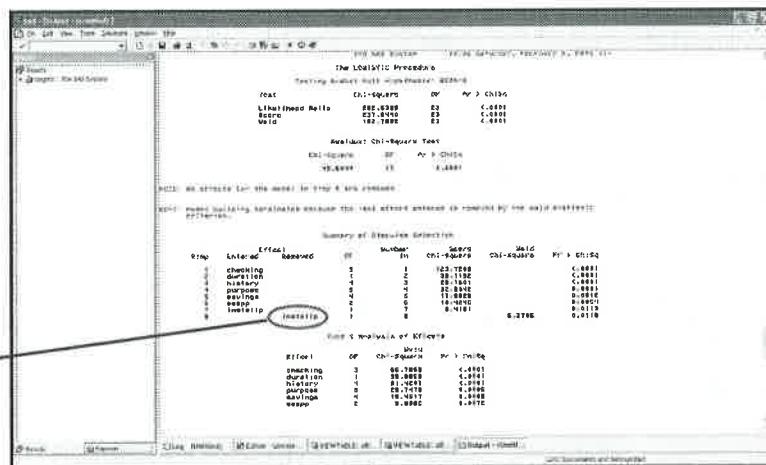
Stepwise Logistic Regression: Example

```
proc logistic data=mydata.applicants;
  class checking history purpose savings
    employed marital coapp resident
    property other housing;
  model good_bad= amount duration age
    installyp checking history purpose
    savings employed marital coapp
    resident property other housing
    /selection=stepwise slentry=0.10
      slstay=0.01;
run;
```

↳ p value

19

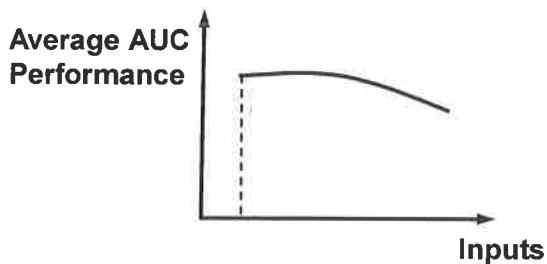
Stepwise Logistic Regression: Example



20

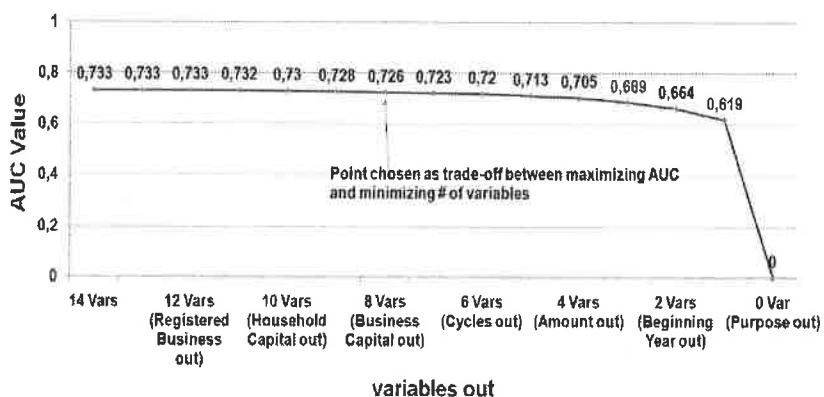
AUC-Based Pruning

1. Start from a model (for example, logistic regression) with all n inputs.
2. Remove each input in turn and reestimate the model.
3. Remove the input giving the best AUC.
4. Repeat this procedure until AUC performance decreases significantly.



21

AUC-Based Pruning: Example



22

VAN GOOL, J., VERBEKE, W., SERCU, P., BAESENS, B., Credit Scoring for Microfinance - is it worth it?, *International Journal of Finance and Economics*, forthcoming, 2011.

Additional Criteria for Input Selection

- **Interpretability of input**

- Do inputs have correct (expected) sign?
 - Interaction effects between inputs might not be considered because of decreased interpretability

- **Computational cost of input**

- How many resources are needed to gather input?
 - E.g., trend variables are typically very important, but also expensive to compute
 - Might opt for a correlated, less predictive, but easier to gather input

- **Legal concerns**

- Cannot use certain inputs, e.g., nationality, ethnic origin, gender, age, ...

23

Chapter 9 Implementing Scorecards

9.1 Implementing Scorecards.....9-3

9.1 Implementing Scorecards

Points-Based Scorecard

Characteristic Name	Attribute	Scorecard Points
AGE 1	Up to 26	100
AGE 2	26 - 35	120
AGE 3	35 - 37	185
AGE 4	37+	225
GENDER 1	Male	90
GENDER 2	Female	180
SALARY 1	Up to 500	120
SALARY 2	501-1000	140
SALARY 3	1001-1500	160
SALARY 4	1501-2000	200
SALARY 5	2001+	240

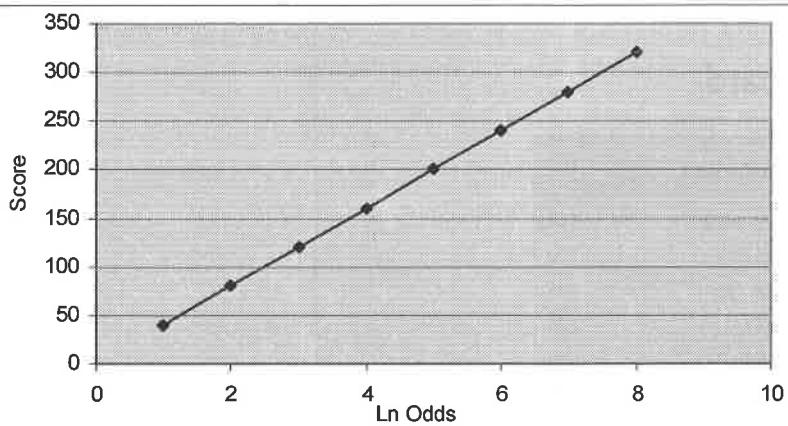
3

Scorecard Scaling

- Refers to the range and format of scores in a scorecard and the rate of change in odds for increases in score
 - Why scale?
 - Implementation software
 - Ease of understanding and interpretation
 - No intercept, logarithmic transformations, ...
 - Continuity with existing scorecards or other scorecards in the company
 - Legal requirements
 - Components
 - Odds at a score
 - Points to double the odds
 - Example: Odds of 50:1 at 600, and odds double every 20 points
-  Performance of scorecard is not affected, only representation.

4

Scorecard Scaling



5

Scorecard Scaling

Example: Odds of 50:1 at 600 and 20 extra points for double odds

$$600 = \ln(50) * \text{factor} + \text{offset}$$

$$620 = \ln(100) * \text{factor} + \text{offset}$$

$$\text{factor} = 20 / \ln(2)$$

$$\text{offset} = 600 - \text{factor} * \ln(50)$$

(Hendrik Wagner)

6

continued...

Scorecard Scaling

$$\begin{aligned}
 score &= \ln(odds) * factor + offset = \\
 &(\sum_{i=1}^n (woe_i * \beta_i) + \beta_0) * factor + offset = \\
 &(\sum_{i=1}^n (woe_i * \beta_i + \frac{\beta_0}{n})) * factor + offset = \\
 &\sum_{i=1}^n ((woe_i * \beta_i + \frac{\beta_0}{n}) * factor + \frac{offset}{n})
 \end{aligned}$$

7

continued...

Scorecard Scaling

The points for each attribute are calculated by multiplying the weight of evidence of the attribute with the regression coefficient of the characteristic, then adding a fraction of the regression intercept, then multiplying by the factor, and finally adding a fraction of the offset:

$$(woe_i * \beta_i + \frac{\beta_0}{n}) * factor + \frac{offset}{n}$$

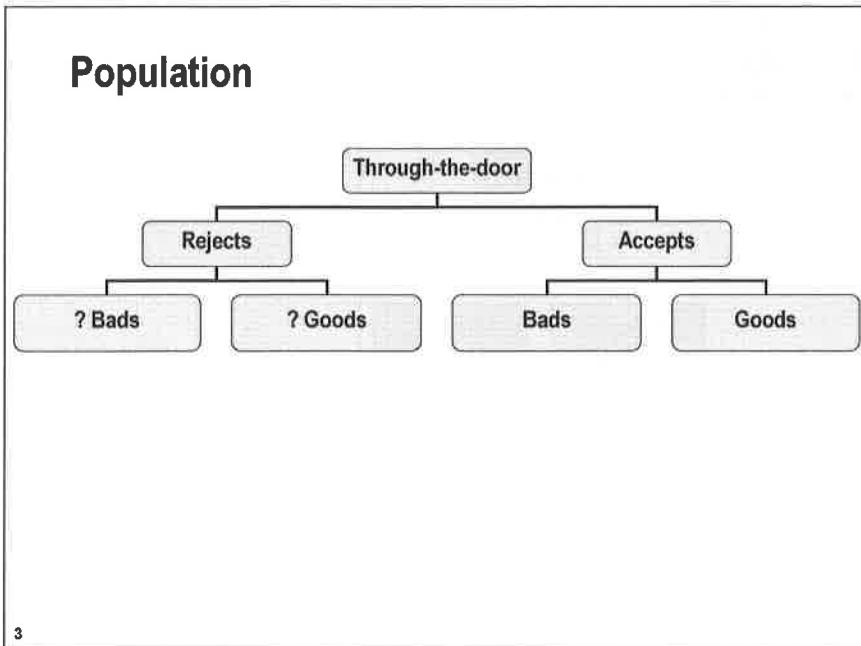
(Hendrik Wagner)

8

Chapter 10 Reject Inference

10.1 Reject Inference.....10-3

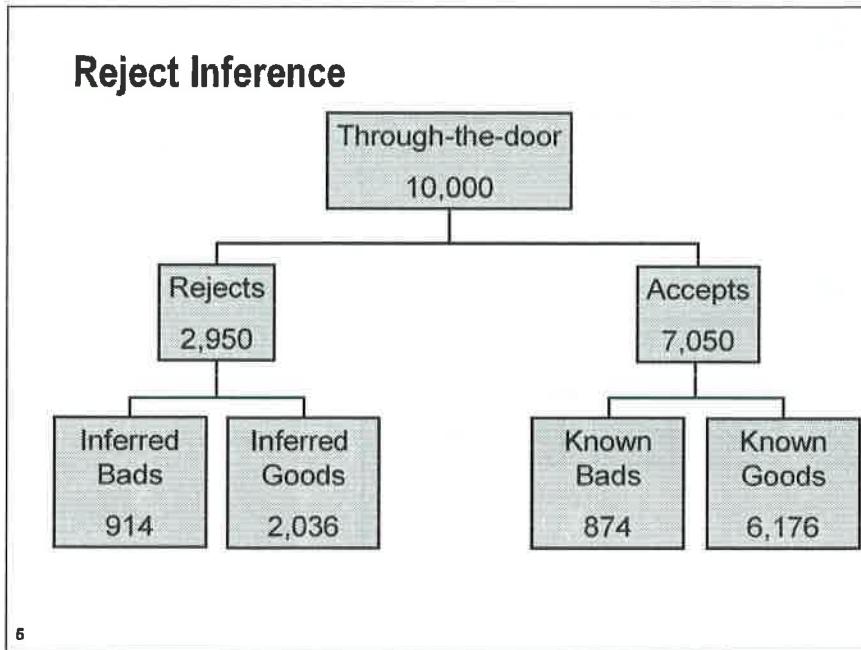
10.1 Reject Inference



Reject Inference

- Problem:
 - Good/Bad information is available only for past accepts, not for past rejects.
 - Reject bias when building classification models.
 - Need to create models for the through-the-door population.
 - If past application policy was random, then reject inference is no problem. However, this is not realistic.
- Solution:
 - *Reject Inference*: A process whereby the performance of previously rejected applications is analyzed to estimate their behavior.

4



Techniques to Perform Reject Inference

- Define as bad
 - Define all rejects as bads
 - Reinforces the credit scoring policy and prejudices of the past
 - In SAS Enterprise Miner (reject inference node):
 - Hard cutoff augmentation
 - Parcelling
 - Fuzzy augmentation
- 6

Hard Cutoff Augmentation

- Also known as Augmentation 1
 - Build a model using known goods and bads
 - Score rejects using this model and establish expected bad rates
 - Set an expected bad rate level above which an account is deemed bad
 - Classify rejects as good and bad based on this level
 - Add to known goods/bads
 - Remodel

P(G)

R ₁	c ₁ 1	"G"
R ₂	c ₁ 3	"B"
R ₃	c ₁ 2	"B"
R ₄	c ₁ 1	"B"

ex 7.5%

build models on accepts + labeled rejects

7

Parcelling

- Rather than classify rejects as good or bad, it assigns them proportional to the expected bad rate at that score.
 - score rejects with good/bad model
 - split rejects into proportional good and bad groups

Score	# Bad	# Good	% Bad	% Good	Reject	Rej - Bad	Rej - Good
0-99	24	10	70.3%	29.7%	348	240	102
100-199	54	196	21.6%	78.4%	654	141	513
200-299	43	331	11.5%	88.5%	345	40	305
300-399	32	510	5.9%	94.1%	471	28	443
400+	29	1,232	2.3%	97.7%	778	18	760

8

continued...

Parcelling

- However:
 - Reject bad proportion cannot be the same as approved.
- Therefore:
 - Allocate higher proportion of bards from reject.
 - Rule of thumb: Bad rate for rejects should be 2 to 4 times that of approved.

9

Fuzzy Augmentation

- Similar to Augmentation 1
- Assigns each reject a partial good and a partial bad class
- Two-stage process
 - Classification
 - Augmentation

10

Fuzzy Augmentation – Classification

- Score rejects with good/bad model.
- Scoring creates $p(\text{good})$ and $p(\text{bad})$ for each reject.
- Don't assign a reject to a class.
- A reject is both good and bad.
- Create two cases from one, $\text{reject}=\text{good}$ and $\text{reject}=\text{bad}$.
- Weigh $\text{reject}=\text{good}$ with $p(\text{good})$.
- Weigh $\text{reject}=\text{bad}$ with $p(\text{bad})$.
- Combine rejects with accepts.
- Remodel.

11

Nearest Neighbor Methods

- Create two sets of clusters: goods and bads.
- Run rejects through both clusters.
- Compare Euclidean distances to assign most likely performance.
- Combine accepts and rejects and remodel.
- However, rejects might be situated in regions far away from the accepts.

12

Additional Techniques to Perform Reject Inference

- Approve all applications
- Three-group approach
- Iterative re-classification
- Mixture distribution modeling

13

Reject Inference

"... There is no unique best method of universal applicability, unless extra information is obtained. That is, the best solution is to obtain more information (perhaps by granting loans to some potential rejects) about those applicants who fall in the reject region."

David Hand, 1998.

Cost of granting credit to rejects versus benefit of better scoring model (for example, work by Jonathan Crook)

14

Reject Inference

- Banasik et al. were in the exceptional situation of being able to observe the repayment behavior of customers who would normally be rejected. They concluded that the scope for improving scorecard performance by including the rejected applicants into the model development process is present but modest.
- Banasik, J., Crook, J.N., Thomas, L.C., "Sample selection bias in credit scoring models," in *Proceedings of the Seventh Conference on Credit Scoring and Credit Control (CSCCVII'2001)*, Edinburgh, Scotland, 2001.

16

Bureau Based Reject Inference

- Bureau data
 - performance of declined applicants on similar products with other companies
 - legal issues
 - difficult to implement in practice – timings, definitions, programming



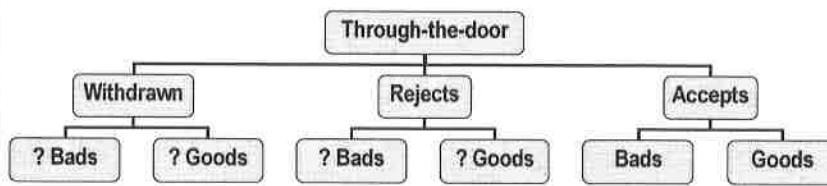
16

Reject Inference

- No method of universal applicability
- Much controversy
- Withdrawal inference
 - Competitive market

17

Withdrawal Inference



18

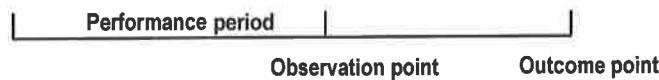
Chapter 11 Behavioral Scoring

11.1 Behavioral Scoring.....11-3

11.1 Behavioral Scoring

Behavioral Scoring

- Performance period versus observation period



- Video clip to snapshot
- Develop classification model (e.g., logistic regression) estimating probability of default during observation period
- Length of performance period and observation period
 - Typically six months to one year

3

Why Use Behavioral Scoring?

- Marketing applications
 - Customer segmentation and targeted mailings
 - Setting credit limits (revolving credit, credit lines)
 - Up-/Down-/Cross-selling
- Debt provisioning
- Authorizing accounts to go in excess
- Collection strategies
 - What preventive actions should be taken if customer starts going into arrears? (early warning system)
- Basel II/Basel III!

4

Using Behavioral Scores for Limit Setting

- When behavioral scores are obtained, they can be used to increase/decrease the credit limits.
- However, the score measures the risk of default given the current operating policy and credit limit.
- So, using the score to change the credit limit invalidates the effectiveness of the score!
- Analogy: "Only those who have no or few accidents when they drive a car at 30 mph in the towns should be allowed to drive at 70 mph on the motorways." (Thomas, Edelman, and Crook 2002)
- Other skills might be needed to drive faster.
- Similarly, other characteristics might be required to manage accounts with large credit limits when compared to accounts with small credit limits.
- Typically, divide the behavioral score into bands and give different credit limit to each band.

5

Variables Used in Behavioral Scoring

- More variables available than in application scoring
 - Application variables
 - Credit bureau data (for example, monthly updates)
 - Repayment and usage behavior characteristics
- Example usage behavior characteristics
 - Maximum and minimum levels of balance, credit turnover, trend in payments, trend in balance, number of missed payments, times exceeded credit limit, times changed home address, ..., during performance period

6

continued...

Variables Used in Behavioral Scoring

- Define derived variables measuring financial solvency of customer.
- Customer can have many products and have different roles for each product.
 - For example, primary owner, secondary owner, guarantor, private versus professional products
- Think about product taxonomy to define customer behavior.
 - For example, average/maximum/minimum/... savings amount, look at different savings products and aggregate
- **The best way to augment the performance of a scorecard is to create better variables!**

7

Aggregation Functions

- Average (sensitive to outliers)
- Median (insensitive to outliers)
- Minimum/Maximum
 - For example, worst status during last 6 months, highest utilization during last 12 months, ...
- Absolute trend $\frac{x_T - x_{T-6}}{6}$
- Relative trend $\frac{x_T - x_{T-6}}{6x_{T-6}}$
- Most recent value, value 1 month ago, ...
- Ratio variable
 - For example: Obligation/income ratio – calculated by adding the monthly house payment and any regular monthly instalment/revolving debt and dividing by the monthly income.
 - Others: current balance/credit limit, ...

8

Impact of Using Aggregation Functions

- Impact of missing values
 - How to calculate trends?
- Ratio variables might have distributions with fat tails (large positive and negative values)
 - Use truncation/winsorizing/capping!
- Data set expands in the number of attributes
 - Input selection!
- What if input selection allows you to choose between, for example, average and most recent value of a ratio?
 - Choose average value for cyclic dependent ratios
 - Choose most recent for structural (non-cyclic) dependent ratios

9

How to Choose the Observation Point

- Seasonality, dependent upon observation point
- Develop 12 PD models
 - Less maintainable
 - Need to backtest/stress test/monitor all these models!
- Use sampling methods to remove seasonality
 - Use three years of data and sample the observation point randomly for each customer during the second year
 - Only one model
 - Gives a seasonally neutral data set
- Migrate from application score to behavioral score using, e.g., a six-month transition period during which a weighted combination of both is used

10

Chapter 12 Defining Default Ratings and Calibrating PD

12.1 Defining Default Ratings and Calibrating PD.....12-3

12.1 Defining Default Ratings and Calibrating PD

Defining Ratings

- Situated at level 2 of the credit risk model architecture.
- A *rating* is a homogeneous pool of obligors that are similar in terms of default risk.
- Ratings are defined because scores are considered too fine granular or too detailed.
- Ratings provide an ordinal measure of credit risk.
- The involvement of a credit expert (or committee) is critical when defining ratings.
- Sometimes a uniform rating scale (aka master rating scale) is adopted across the entire firm.
- Approaches
 - Map onto rating agency scale
 - Statistical approaches

3

Map onto Agency Rating Scale

Define internal ratings so as to mimick as closely as possible the one-year default rates reported by the rating agencies.

LTRating	Moenly %	Fitch	S&P	Average
AAA	0.00%	0.00%	0.00%	0.00%
AA+	0.00%	0.00%	0.00%	0.00%
AA	0.00%	0.00%	0.00%	0.00%
AA-	0.02%	0.00%	0.02%	0.01%
A+	0.00%	0.00%	0.05%	0.02%
A	0.03%	0.00%	0.04%	0.02%
A-	0.04%	0.12%	0.04%	0.07%
BBB+	0.17%	0.29%	0.21%	0.22%
BBB	0.16%	0.13%	0.30%	0.20%
BBB-	0.34%	0.56%	0.39%	0.43%
BB+	0.75%	0.91%	0.65%	0.77%
BB	0.78%	1.80%	0.96%	1.18%
BB-	2.07%	1.94%	1.80%	1.94%
B+	3.22%	1.82%	0.21%	2.75%
B	5.46%	1.78%	8.87%	5.37%
B-	10.46%	1.69%	12.99%	8.38%
CCC	20.98%	26.07%	31.66%	26.04%
Inv.Gr.	0.08%	0.11%	0.11%	
Spec.Gr.	5.15%	3.27%	4.65%	
All	1.74%	0.65%	1.61%	

LTRating	Default rate	
	1-year	5-years
Aaa	0.00%	0.09%
Aa	0.01%	0.20%
A	0.02%	0.56%
Baa	0.21%	2.25%
Ba	1.31%	11.85%
B	5.69%	29.73%
Caa	20.98%	57.01%
Inv.Gr.	0.08%	0.93%
Spec.Gr.	5.15%	23.49%
All	1.74%	7.73%

4

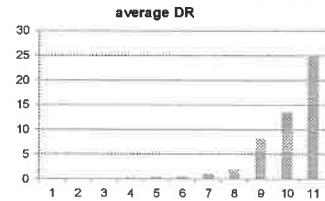
Statistical Approaches for Defining Ratings

Maximize the following objective function for K ratings:

$$\sum_{i=1}^{K-1} \frac{\sqrt{|DR_i - DR_{i+1}|}}{\sqrt{\frac{DR_i(1-DR_i)}{n_i} + \frac{DR_{i+1}(1-DR_{i+1})}{n_{i+1}}}}$$

Define ratings such that the default rates double per rating decrease.

- Default rates increase exponentially for decreasing ratings



5

Rating Migration Analysis

- Study migrations between ratings during a specific time period (for example, 12 months)
- Represent as rating migration matrix M as follows:

From/to	AAA-AA	A	BBB	BB	B	CCC	C	D
AAA-AA	91.30%	5.62%	1.11%	1.03%	0.84%	0.07%	0.02%	0.01%
A	5.98%	85.90%	5.71%	1.67%	0.53%	0.09%	0.09%	0.03%
BBB	0.66%	7.02%	84.31%	6.96%	0.78%	0.12%	0.10%	0.05%
BB	0.08%	0.58%	3.99%	89.28%	4.81%	0.57%	0.43%	0.26%
B	0.08%	0.12%	0.26%	10.95%	84.07%	1.86%	1.60%	1.06%
CCC	0.00%	0.09%	0.18%	1.99%	15.10%	63.47%	10.04%	9.13%
C	0.10%	0.10%	0.10%	1.40%	1.40%	4.60%	74.58%	17.72%
D	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	100%

6

continued...

Rating Migration Analysis

- All transition probabilities are between 0 and 1
- Sum across the rows is always 1
- When smooth transitions are desired, might require transition probabilities to decrease when getting further away from the diagonal
- Many migration matrices are typically diagonally dominant, implying that many obligors maintain their rating and only a minority migrates
- Can distinguish between upgrades (for example, going from BBB to A) and downgrades (for example, going from A to BBB)

7

continued...

Rating Migration Analysis

- Can think of rating migration as a Markov process whereby state/rating at time $t+1$ only depends upon state at time t
- Default state is absorbing state
- E.g., suppose we start with 1000 obligors in rating AAA-AA
- Can find state after 12 months using matrix multiplication, P.M, with $P=[1000 \ 0 \ 0 \ 0 \ 0 \ 0 \ 0]$ and M migration matrix as follows:

$$\begin{bmatrix} 1000 & 0 & 0 & 0 & 0 & 0 & 0 \end{bmatrix} \times \begin{bmatrix} 91.30\% & 5.62\% & 1.11\% & 1.03\% & 0.84\% & 0.07\% & 0.02\% & 0.01\% \\ 5.98\% & 85.90\% & 5.71\% & 1.67\% & 0.53\% & 0.09\% & 0.09\% & 0.03\% \\ 0.66\% & 7.02\% & 84.31\% & 6.96\% & 0.78\% & 0.12\% & 0.10\% & 0.05\% \\ 0.08\% & 0.58\% & 3.99\% & 89.28\% & 4.81\% & 0.57\% & 0.43\% & 0.26\% \\ 0.08\% & 0.12\% & 0.26\% & 10.95\% & 84.07\% & 1.86\% & 1.60\% & 1.06\% \\ 0.00\% & 0.09\% & 0.18\% & 1.99\% & 15.10\% & 63.47\% & 10.04\% & 9.13\% \\ 0.10\% & 0.10\% & 0.10\% & 1.40\% & 1.40\% & 4.60\% & 74.58\% & 17.72\% \end{bmatrix} = [913 \ 56 \ 11 \ 10 \ 8 \ 1 \ 1 \ 0]$$

8

continued...

Rating Migration Analysis

State after 24 months can then be calculated as P.M.M

Can get multiperiod transition probabilities by multiplying transition matrix by itself

- For example, n-period transition matrix= M^n
- Can be extended using, for example, mover-stayer model whereby only a fraction of obligors (movers) change rating and others keep rating (stayers)
- Markovian property sometimes questioned as empirical evidence seems to suggest the following:
 - Downgrades tend to be more easily followed by further downgrades (autocorrelation over time)
 - Duration dependence effect: the longer an obligor keeps the same rating, the lower the migration probability
 - Migration probabilities tend to be correlated with business cycle

9

Rating Philosophy

- Before validation starts, one needs to answer the question: "What is the rating system supposed to do?"
- "Rating philosophy is the term used to describe the assignment horizon of a borrower rating system." (FSA, CP 05/03, par. 7.66)
- Rating philosophy should be clearly articulated in bank rating policy.
- Federal Register, wholesale exposures
 - A bank's rating policy must describe its ratings philosophy and how quickly obligors are expected to migrate from one rating to another in response to economic cycles.
 - Analyze rating philosophy by means of migration analysis.

10

Rating Philosophy

- Point-in-Time (PIT) Ratings
 - Take into account obligors specific, cyclical, and non-cyclical information
 - Rating changes rapidly with macroeconomic situation (lots of rating mobility)
 - PD is the best estimate of the obligor's default during next 12 months
- Through-the-Cycle (TTC) Ratings
 - Take into account non-cyclical information
 - Obligors with same TTC rating share similar stressed PDs
 - Rating robust with respect to macro-economic situation (not much rating mobility)
 - PD is best estimate of the obligor's default during a credit cycle
- Many hybrids exist; PIT and TTC represent two ends of a continuum!

11

Rating Mobility

- Most extreme example of a TTC rating system is the identity matrix (no rating changes).
- The higher the concentration on the diagonal, the lower the mobility.
- The higher the mobility metric, the more PIT the risk rating system.
- Calculate mobility using, e.g., the following metrics:

$$M_{L1}(P) = \frac{\sum_{i=1}^N \sum_{j=1}^N |P_{ij} - I_{ij}|}{N^2}$$

$$M_{L2}(P) = \sqrt{\frac{\sum_{i=1}^N \sum_{j=1}^N (P_{ij} - I_{ij})^2}{N^2}}$$

- By computing these metrics, we can compare internal rating philosophy against external benchmarks (e.g., S&P's transition matrix).

12 Measurement, Estimation and Comparison of Credit Migration Matrices,
Schuermann and Jafry, Journal of Banking and Finance 2004

PD Calibration

- Calculate empirical one-year realized default rates for borrowers in a specific grade/pool

$$\text{1-year default rate for rating grade } X = \frac{\text{Number of obligors with rating } X \text{ at the beginning of the given time period that defaulted during the time period}}{\text{Number of obligors with rating } X \text{ at the beginning of the given time period}}$$

- Numerator can also measure number of default events (which will be higher than the number of defaulted obligors if they can default (and recover) multiple times in the same period)
 - Make sure to also appropriately count in LGD!

13

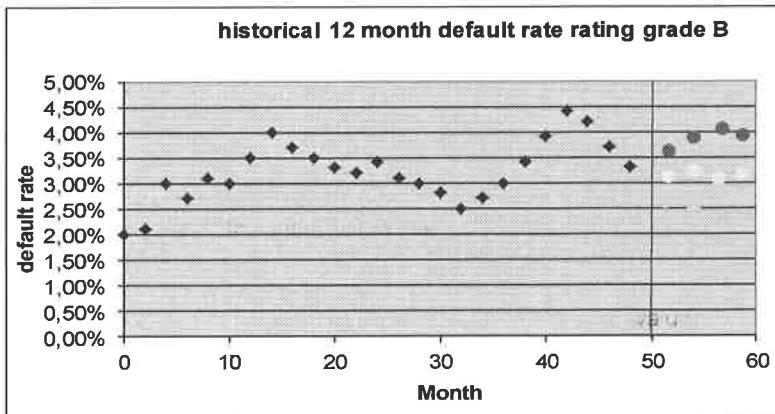
continued...

PD Calibration

- Use five years of default rates to estimate future PD (retail)
- How to calibrate the PD?
 - Average
 - Maximum
 - Upper percentile (stressed/conservative PD)
 - Dynamic model (time series, for example, ARIMA)
 - Requires sufficient historical data
 - Can link with macro-economic conditions (for example, GDP)
 - For example, $DR_t = \beta_0 + \beta_1 GDP_{t-1}$
- Make sure to make forward-looking adjustments!

14

PD Calibration



16

Chapter 13 LGD Modeling

13.1 LGD Modeling 13-3

13.1 LGD Modeling

Overview

- Level 0: data
 - Definition of default
 - Definition of LGD
 - Ways of measuring LGD
 - Preprocessing data for LGD
- Level 1: predictive model
 - Predictive model for ordinal ranking of LGD
 - Segmentation/Regression
 - Two-stage models
- Level 2: Ratings and calibration
 - Defining LGD ratings
 - (Economic downturn) calibration
- PD/LGD correlation

3

Level 0: Definition of Default

- Moody's, S&P, and Fitch use different definitions of default
 - Hence, default numbers and LGD distributions are hard to compare
- Default types:
 - **Operational default:** due to technical issues at counterpart side (e.g., customer is accidentally late)
 - **Technical default:** due to internal information system issue (e.g., bank expected payment on another account in another system)
 - **Real default:** due to financial problems
- In case of default:
 - Cure: return to performing status
 - Restructuring/Settlement: recovery plan
 - Liquidation: assets are sold/bankruptcy procedure started
- LGD values depend on default definition (correlation between default definition, default rates and LGD values)
- Default definition has impact on cure rates and as such also on bi-modal LGD distribution!

4

Level 0: Definition of LGD

- LGD is the final economic loss of an account as a percentage of the exposure, given that the account goes in arrears (=1-recovery rate)
- Economic loss versus Accounting loss
- Problems:
 - Extra Costs
 - Realizing collateral value (haircuts), administrative costs (letters, telephone calls), legal costs, time delays in what is recovered (economic loss!), ...
 - Extra Benefits
 - Interest on arrears, penalties for delays, commissions, ...
- Issue weighted LGD versus issuer weighted LGD
 - If cash flows cannot be allocated to a specific issue, use issuer-weighted LGD

5

Level 0: Ways of Measuring LGD

Workout LGD (corporate/consumer)

- Discount cash flows from workout/collections process.
- Direct and indirect costs (e.g., operating costs of workout department)
- **Most often used!**

Market LGD (corporate)

- Use market price of defaulted bonds/loans one month after default
- Only for debt securities that trade in the market
- E.g., Moody's LossCalc

6

continued...

Level 0: Ways of Measuring LGD

Implied historical LGD

- Use PD estimate and observed total losses to derive implied LGD
- Based on total losses, estimate EL and then $LGD = EL/PD$
- Often though historical losses are accounting losses!
- Only allowed for retail exposure class! (par. 239, CEBS, CP10)

Implied market LGD (corporate)

- Use market price of risky but not defaulted bond prices using the asset pricing models (structural/reduced form)
- Spread above risk-free rate reflects EL
- But: market price is only partially determined by credit risk!
- Not often used because difficult!

Both market and workout LGD discussed in Federal Register

7

Level 0: LGD According to Basel II

- “The definition of loss used in estimating LGD is **economic loss**.” (par. 460)
- “Owing to the potential of very long-run cycles in house prices which short-term data may not adequately capture, during this transition period, LGDs for retail exposures secured by **residential properties cannot be set below 10%** for any sub-segment of exposures ...” (par. 266)
 - Floor also applicable in U.S.
- Foundation approach
 - For Corporates/Sovereigns/Banks
 - Senior claims on corporates, sovereigns and banks not secured by recognised collateral will be assigned a 45% LGD.” (par. 287)
 - “All subordinated claims on corporates, sovereigns, and banks will be assigned a 75% LGD.” (par. 288)

8

Level 0: Constructing an LGD Data Set

- See Federal Register or paragraph 223, CEBS, CP10
- Data set should cover at least a complete business cycle
- Default definition and cures
- Decide on workout period
- Incomplete workouts
- Discount factor
- Negative LGDs versus LGDs>100%
- Indirect costs
- Drivers of LGD

9

Level 0: Complete Business Cycle

- For the retail portfolio
 - Minimum 5 years
- For wholesale (corporates, sovereigns, banks)
 - Minimum 7 years
- Need not attach equal importance to each year of data
- See Federal Register

10

Level 0: Default Definition and Cures

- All defaults should be included
- Same default definition as for PD
- Cures
 - $LGD=0$
 - Depends strongly on default definition (cf. supra)!
 - Relaxing default definition typically increases cures
- Multiple defaults
 - Only include last default event
 - PD and EAD also related to this

11

Level 0: Length of Workout Period

- Non-recovered value is < 5% of EAD (BIS, HKMA)
- One year after default (BIS, HKMA)
- Time of repossession (HKMA)
- Time of selling off the debt to a collection agency (for example, United Kingdom)
- 2, 3, 4, 5 years?
- Combination

12

Level 0: Incomplete Workouts

- “*The calculation of default-weighted average of realised LGDs, requires the use of all observed defaults in the data sources. Observed defaults include incomplete work-out cases, although they will not have values for the final realisation of LGD because the recovery process has not ended.*” (CEBS, CP10, par. 231, 2005)
- “*Institutions should incorporate the results of incomplete workouts (as data/information) into their LGD estimates, unless they can demonstrate that the incomplete workouts are not relevant.*” (CEBS, CP10, par. 231, 2005)

13

Level 0: Incomplete Workouts

- No regulatory prescription as to how incomplete workouts should be treated!
- Options:
 - Ignore incomplete workouts when not relevant
 - Look at current LGD of an incomplete workout and use this for estimation (taking into account margin for future expected costs)
 - Very conservative!
 - Build predictive models estimating final LGD based upon what has been observed already (% collected, time of collection, ...)
 - Survival analysis models whereby recovery amount is considered as censored variable at e.g. 12 months
 - See, e.g., Stoyanov S., Application LGD Model Development, *Credit Scoring and Credit Control XI Conference*, 2009.

14

Level 0: Discount Rate

- Ongoing debate between supervisors and firms
- Common industry approaches
 - Use contractual rate at time of default
 - Use risk free rate + a risk premium (e.g., 1%)
 - Use 4%
 - It might be useful to perform a sensitivity analysis (effect could be quite limited anyway)!
- Subject to discussion (for example, FSA Expert group on LGD)
 - The EG agrees that the use of the contractual rate as the discount rate is conceptually inappropriate.
 - The group proposes that the discount rate for this asset class should be close to the risk-free rate, so long as firms can evidence and justify sufficient conservatism in their estimation of the downturn. One potential approach to a discount rate for this asset class could be the risk-free rate plus an appropriate premium.

15

Level 0: Discount Rate

“The EG seeks assurance from the supervisor that the UK industry would not be required to adopt overly theoretical approaches which may hinder the evolution of methodologies and their application.”

(FSA, October 2005)

16

Level 0: Negative LGDs and LGDs > 100%

- Negative LGDs
 - Recovery rate > 100%
 - Reasons
 - EAD measured at moment of default whereas claim on the borrower increases afterwards (fines, fees, ...) and everything is recovered
 - Result of gain in collateral sales
 - Truncate negative LGDs to 0
- LGDs > 100%
 - Recovery rate < 0%
 - Additional costs incurred but nothing recovered
 - Also because of definition of EAD; additional drawings after time of default considered part of LGD

17

continued...

Level 0: Negative LGDs and LGDs > 100%

- “Reference data sets may contain individual loss observations that are less than 0 percent or greater than 100%.” (Federal Register)
 - Banks are not required to truncate to 0 and 100
 - However, final LGD estimates should not be negative or zero!

18

Level 0: Indirect Costs

- “The definition of loss used in estimating LGD is economic loss...This must include material discount effects and material direct and indirect costs associated with collecting on the exposure.” (paragraph 460 of the Accord)
- “Work-out and collection costs should include the costs of running the institution’s collection and work-out department, the costs of outsourced services, and an appropriate percentage of other ongoing costs, such as corporate overhead.” (par. 205, CEBS, CP10)
- “Cost data comprise the material direct and indirect costs associated with workouts and collections.” (Federal Register)
- Material indirect costs, costs of running the collection and workout department, costs of outsourced services, appropriate percentage of overhead, must be included (Federal Register).

19

Level 0: Allocating Workout Costs

Calculate a cost rate: average (e.g., between 0.5%-3% of EAD) or via accounting exercise

Example

Year	Total EAD of files in workout (end of year)	Internal workout costs per year	Amount recovered during year
2007	1000	20	250
2008	1500	28	500
2009	800	12	240
2010	1250	27	350

Option 1: Use EAD at time of default as denominator

- Assumption: higher costs for higher exposures at default
- Time weighted=1/4*[20/1000+28/1500+12/800+27/1250]=1.8%
- Pooled=[20+28+12+27]/[1000+1500+800+1250]=1.91%
- Disadvantage: for an individual file, the cost rate has to be multiplied by the number of years the workout lasted

20

Level 0: Allocating Workout Costs

Option 2: Use the amount recovered as denominator

- Assumption: higher costs for higher recoveries
- Time weighted
 $=1/4*[20/250+28/500+12/240+27/350]=6.5\%$
 Pooled
 $=[20+28+12+27]/[250+500+240+350]=6.49\%$
- Advantage is that this is independent of the length of the workout process because each amount was recovered during one year only!
- Much simpler to implement!

21

Level 0: Calculating Individual LGD

Recovery rate=net actualized cash flow (NCF) /EAD

Option 1

- c =direct + indirect cost rate (as a % of amount outstanding, taking into account length of workout period)

$$NCF = \frac{\sum_{t=1}^n CF_t}{(1+i)^t} - c * n * EAD$$

Option 2

- c =direct + indirect cost rate (as a % of recovered amount)
- n =number of years of recovery period

$$NCF = \frac{\sum_{t=1}^n CF_t * (1-c)}{(1+i)^t}$$

22

Level 0: Drivers for Predictive Modeling of LGD

■ Borrower characteristics

- Creditworthiness (PD, rating, application/behavioral score, FICO score, trends in creditworthiness, rating changes, ...)
- Marital status, gender (?), salary, time at address, time at job, ...
- Industry sector, sector indicators
- Size of the company
- Legal form of the company
- Age of the company
- Intensity of relationship (number of years client, number of products, ...)
- Balance sheet information (revenue, total assets, solvency/profitability/liquidity ratios, ...)

23

continued...

Level 0: Drivers for Predictive Modeling of LGD

■ Macro-economic factors

- GDP (growth)
- Default rates
- Inflation, unemployment rate
- Interest rate (has impact on discount factor!)

24

continued...

Level 0: Drivers for Predictive Modeling of LGD

- **Loan Characteristics**

- Type and value of collateral (e.g., real estate, cash, inventories, guarantees, ...)
- Real estate: flat, apartment, villa, detached/semi-detached house, ...
- Loan to Value (LTV)
 - ratio of the value of the loan (=exposure) to the value of the underlying asset; LTV at start versus LTV at repossession
 - Estimate and apply haircuts reflecting market value!
 - Haircut=(current market value-forced sales price)/current market value
 - Current market value=current HPI/start HPI* start valuation
 - Define LTV through time (compute trends, ...)

25

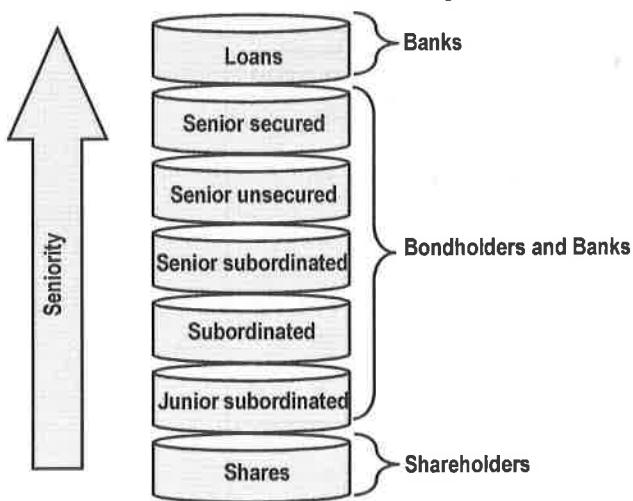
continued...

Level 0: Drivers for Predictive Modeling of LGD

- EAD
- Debt seniority
 - Absolute debt seniority: senior secured, subordinated, ...
 - Relative debt seniority: debt amount above/total debt; debt amount below/total debt
- Remaining maturity
- Country-related features
 - Geographical region (e.g., ZIP code)
 - For example, how creditor-friendly is the bankruptcy regime?

26

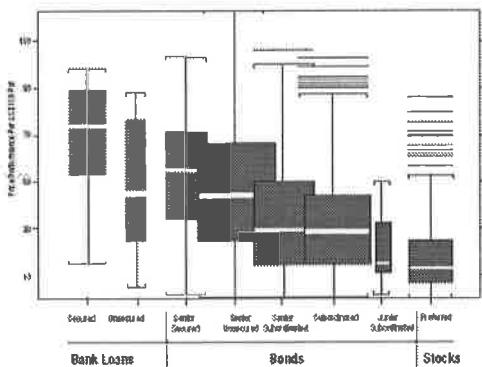
Level 0: LGD Drivers: Seniority



27

Level 0: LGD Drivers: Debt Type and Seniority

Default Recovery by Debt Type and Seniority, 1981-2000



This figure is adapted from Moody's 2001 annual default study; see Exhibit #20 in Hamilton, Gupton & Berthault [2001]. It highlights the wide variability of recoveries even within individual seniority classes. The shaded boxes cover the inter-quartile range with the median marked as a white horizontal line. Squared brackets cover the data range except for outliers that are marked as horizontal lines.

28

Level 0: LGD Drivers: Industry Impact

Industry	Avg. Recovery (cents on dollar)	Industry	Avg. Recovery (cents on dollar)
Utilities	74	High Technology / Office Equipment	47
Insurance & Real Estate	37	Aerospace / Auto / Capital Goods	52
Telecommunications	53	Forest, Building Products / Homebuilders	54
Transportation	39	Consumer / Service	47
Financial Institutions	59	Leisure Time / Media	52
Healthcare / Chemicals	56	Energy & Natural Resources	60

Acharya, Bharath, Srinivasan, 2003

29

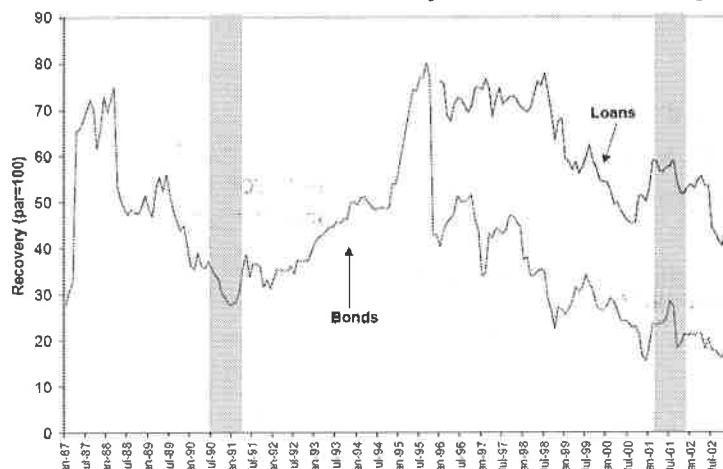
Level 0: LGD Drivers: Firm Impact

Table 3 LGD by Business Unit JPMC Resolved Defaults (1Q82-4Q99)						
Business Units	Obligor Count	Average Time-to-Resolution (Years)	Net Charge-Offs		Discounted LGD	
			Mean	Standard Deviation	Mean	Standard Deviation
Large Corporates (U.S.)	676	3.33	23.8%	34.2%	41.6%	30.9%
Large Corporates (non-U.S.)	268	2.58	22.9%	33.8%	37.3%	33.2%
Real Estate	719	2.23	29.8%	36.6%	42.0%	33.7%
Emerging Markets	394	3.04	25.8%	39.5%	42.2%	35.6%
Middle Market	1,264	2.15	30.0%	40.4%	40.3%	38.4%
Private Banking	310	1.66	25.4%	40.9%	34.5%	38.3%
Total	3,761	2.43	27.0%	37.9%	39.8%	35.4%

Measuring LGD on Commercial Loans: an 18-year Internal Study,
The RMA Journal, May 2004

30

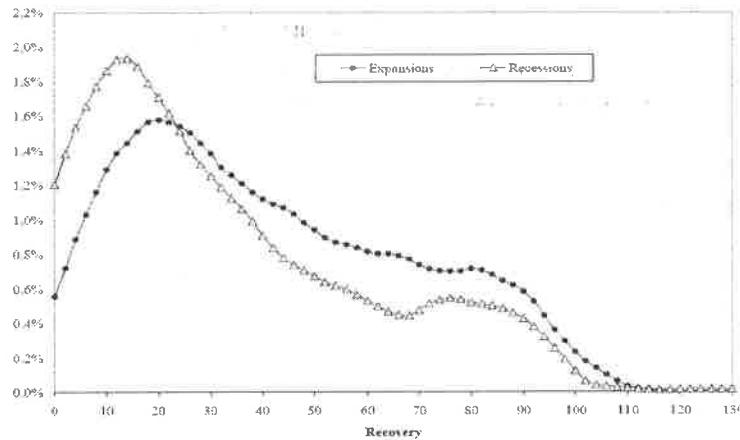
Level 0: LGD Drivers: Impact of Economy



Altman data set on U.S. firms. Shaded regions recessions.

31

Level 0: LGD Drivers: Impact of Economy(1970–2003)



Scheurmann 2005

32

Level 0: Data Preprocessing

- Missing values
 - Keep/Delete/Impute
- Outlier detection and treatment
 - Truncation procedures (cf. supra)
- Coarse classification\categorization
 - E.g., industry sector
 - Decision trees (CART) or chi-squared analysis!

33

Level 1: Challenges in LGD Modeling

- Definition of LGD
- Dependent variable continuous between 0% and 100%
- “The CRD does not permit the use of estimates based purely on judgmental considerations.”
(par. 228, CEBS, CP10)
- Purpose at level 1 is to provide ordinal ranking of LGD!
 - High scores for high losses, low scores for low losses

34

Level 1: LGD Modeling Approaches Observed in Industry

One-stage

- Segmentation
 - Expert based (e.g., based on experience)
 - Statistical: regression trees (e.g., CART)
- Regression
 - Linear regression
 - Linear regression with beta transformation
 - Logistic regression
 - Multinomial regression

Two-stage models

35

Level 1: Segmentation

- Use historical averages or expert opinions to estimate LGD
- Table look-up
- Segmented per
 - Debt type
 - Seniority class (senior versus subordinated)
 - Collateral type (for example, secured versus unsecured)
 - Loan purpose
 - Business segment
- However, one typically observes a wide variability of loss rates per segment, making averages less suitable to work with.

36

Level 1: Example Segmentation Approach

LGD (percentage of EAD lost in default)			
	Low coverage	Medium coverage	High coverage
Collateral Type 1	25	12	5
Collateral Type 2	5	3	1
Collateral Type 3	16	16	14
Collateral Type 4	90	40	25

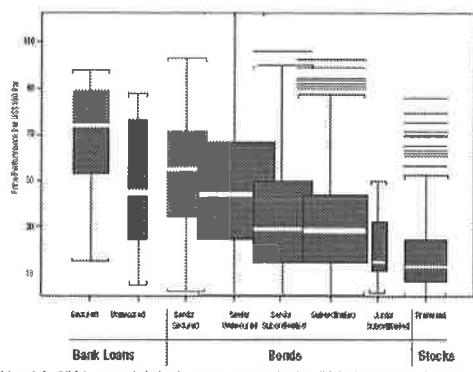
Collateral, for example, real estate, inventories, cash, ...
 Collateral coverage (related to LTV)

Levonian 2006

37

Level 1: Segmentation Approach to LGD Modeling

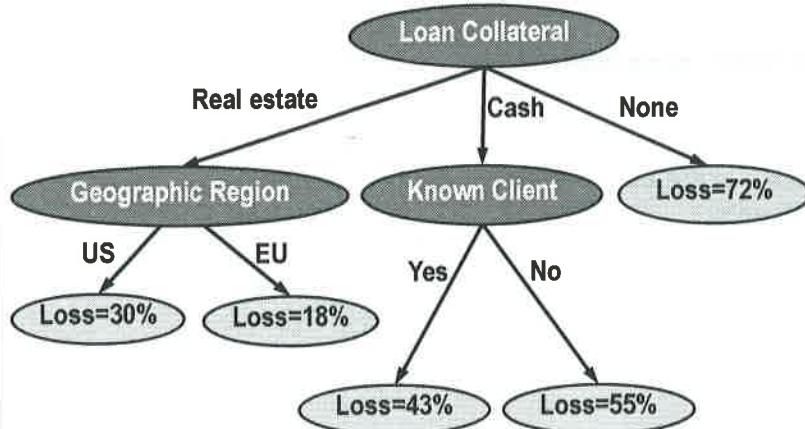
Default Recovery by Debt Type and Seniority, 1981-2000



This figure is adapted from Moody's 2001 annual default study; see Exhibit #20 in Hamilton, Gupta & Berthault (2001). It highlights the wide variability of recoveries even within individual seniority classes. The shaded boxes cover the inter-quartile range with the median marked as a white horizontal line. Squared brackets cover the data range except for outliers that are marked as horizontal lines.

38

Level 1: Regression Trees (e.g., CART)



39

Regression Trees

Splitting Decision

- Mean Squared Error (variance in SAS Enterprise Miner)

$$I(N) = \frac{1}{N} \sum_{i=1}^N (Y_i - \bar{Y})^2, \text{ with } N = \text{number of observations in node}$$

- ANOVA/F-test (probF in SAS Enterprise Miner)

$$F = \frac{SS_{between}/(B-1)}{SS_{within}/(N-B)} \sim F_{N-B, B-1}$$

$$SS_{between} = \sum_{b=1}^B N_b (\bar{Y}_b - \bar{Y})^2,$$

with B = number of branches, N_b number of observations in branch b , \bar{Y}_b the mean in branch b

$$SS_{within} = \sum_{b=1}^B \sum_{i=1}^{N_b} (Y_{bi} - \bar{Y}_b)^2, \text{ with } \bar{Y}_b \text{ the mean in branch } b$$

Compute the p-value. Low p-value indicates good split!

40

Regression Trees

Stopping decision

- Use mean squared error (MSE) measured on the validation set

Assignment decision

- Average of the target values in the training data in the leaf (+ confidence levels)

41

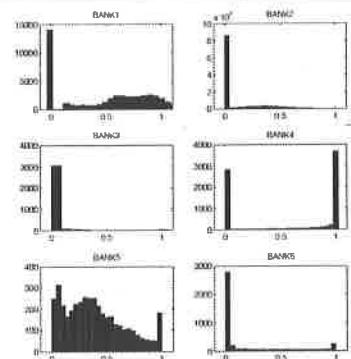
Level 1: Regression LGD Modeling

- Model recovery rate or LGD as a linear function of loss drivers
 - $LGD = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n$
- Problem:
 - Linear regression output not bounded between 0 and 1
 - Can be remedied by using sigmoid transformation
 - $LGD = 1/(1 + \exp(-(\beta_0 + \beta_1 x_1 + \dots + \beta_n x_n)))$
 - Recovery rate is not normal distribution
 - Can be unimodal or even bimodal
 - Regression assumes errors are normal
- Solution
 - Map a Beta distribution onto the LGD data
 - Beta distributions are very flexible and versatile
 - If LGD distribution is unimodal: 1 Beta distribution might suffice
 - If LGD distribution is bimodal: might need 2 (or more) Beta distributions (Beta mixture models)

42

Level 1: Example LGD Distributions

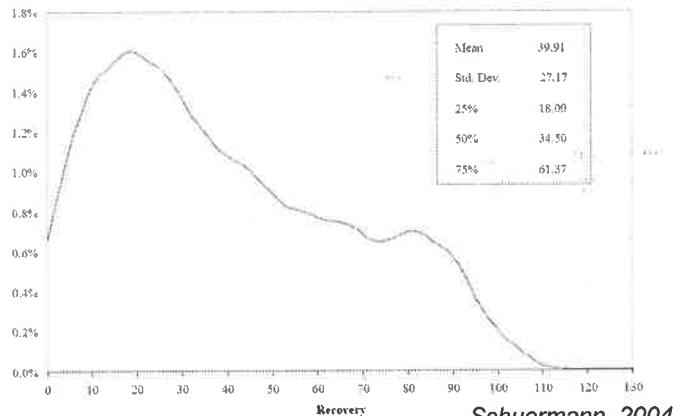
Dataset	Type	Inputs	Total size	Training size	Test size
BANK1	Personal loans	44	47 853	31 905	15 948
BANK2	Mortgage loans	18	119 211	71 479	39 712
BANK3	Mortgage loans	14	3 351	2 232	1 119
BANK4	Revolving credit	12	7 889	5 260	2 629
BANK5	Mortgage loans	35	4 057	2 733	1 364
BANK6	Corporate loans	21	4 276	2 851	1 425



Loterman G., Brown I., Martens D., Mues C., Baesens B., Benchmarking regression algorithms for loss given default modeling, *International Journal of Forecasting*, forthcoming, 2011.

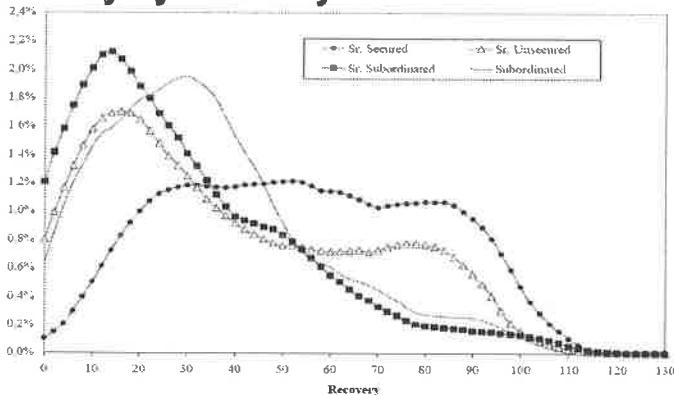
43

Level 1: Bimodal Distribution of Recovery Rate: All Bonds and Loans (Moody's 1970–2003)



44

Level 1: Probability Density of Recovery by Seniority



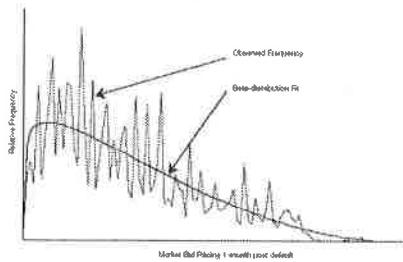
This is market LGD, and other patterns have been observed for workout LGD!

45

Schuermann 2004

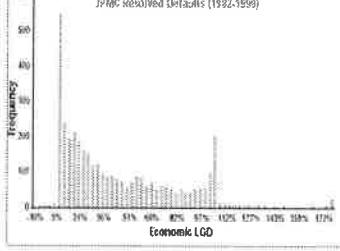
Level 1: LGD Regression Modeling

Beta-distribution Fit to Recoveries



Taken from LossCalc: model for predicting loss given default (LGD), Moody's KMV, February 2002

Figure 1 Distribution of Economic LGD for the Wholesale Bank JPMC Resolved Defaults (1982-1999)



Measuring LGD on Commercial Loans: an 18-year Internal Study, The RMA Journal, May 2004 (JPMorgan Chase)

46

Level 1: The Beta Distribution, Mean μ , Variance σ^2

$$f(x) = \frac{\Gamma(\alpha + \beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1} (1-x)^{\beta-1}, \text{ with } \Gamma(x) \text{ Euler's gamma function}$$

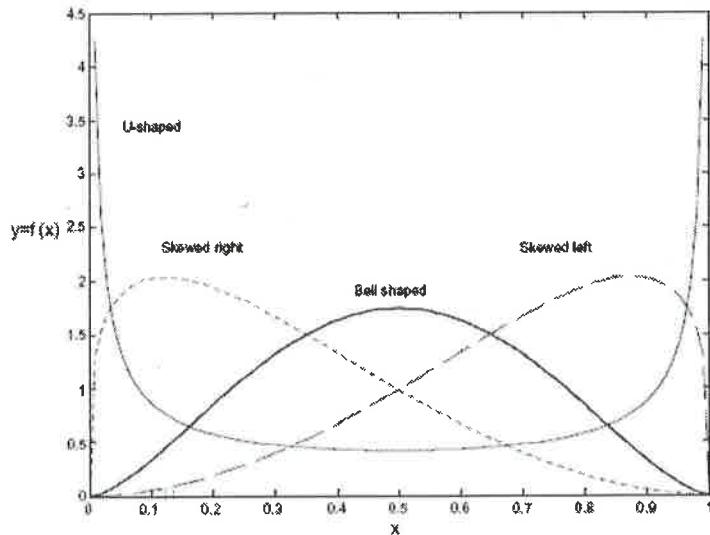
$$\mu = \frac{\alpha}{\alpha + \beta} \text{ and } \sigma = \sqrt{\frac{\alpha\beta}{(\alpha + \beta)^2(1 + \alpha + \beta)}}$$

- Bounded between 0 and 1! (or any upper bound you like)
- Center parameter α and shape parameter β allow to model a wide variety of distributions
- α and β can be estimated using maximum likelihood or the method of moments

$$\alpha = \left[\mu^2 \frac{(1-\mu)}{\sigma^2} \right] - \mu \quad \text{and} \quad \beta = \alpha \left(\frac{1}{\mu} - 1 \right)$$

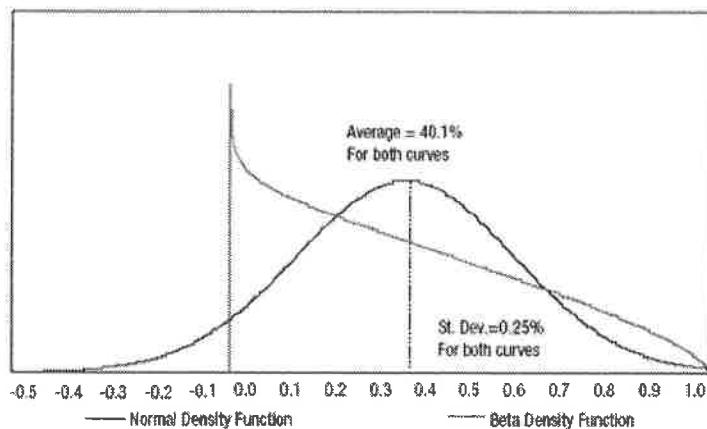
47

The Beta Distribution



48

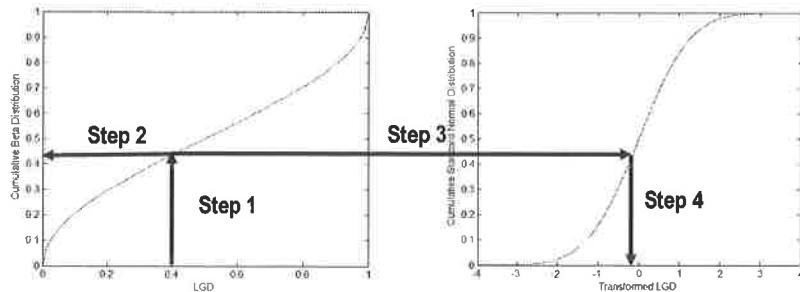
Level 1: Beta versus Normal Distribution



Taken from LossCalc: model for predicting loss given default (LGD),
Moody's KMV, February 2002

49

Level 1: Transforming a Beta Distributed Variable to a Standard Normal Distributed Variable



50

Level 1: Modeling LGD Using Linear Regression

- Step 1: estimate α and β using the method of moments
- Step 2: transform the raw LGDs using the cumulative Beta distribution
- Step 3: transform the numbers obtained in step 2 using the inverse standard normal distribution
- Step 4: perform linear regression using the numbers obtained in step 3

Go the other way around to get
LGD predictions based on the regression!

51

Level 1: Logistic Regression

See, e.g., Stoyanov, S., Application LGD Model Development, Credit Scoring and Credit Control XI Conference, 2009.

Transform continuous LGD variable to binary variable by generating a random number: if LGD > random number then LGD Binary=1 (Bads); else LGD Binary =0 (Goods)

LGD	Random Number	LGD Binary
0.88	0.24	1
0.05	0.12	0
0.22	0.90	0
0.76	0.62	1

Can also use manual cutoff (e.g. 0.2) to create binary target
 52 Build logistic regression model predicting LGD Binary

Level 1: Ordinal Classification

- LGD values are binned from low to high.
 - LGD1=[0%;5%], LGD2=[5%;10%], ...LGD20=[95%;100%]
 - Or, use recovery ratings from rating agencies and perform mapping.
- Use the cumulative logistic regression model to perform the mapping:

$$P(C \leq i | \mathbf{x}) = \frac{1}{1 + e^{-\theta_i + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n}}$$

$$P(C = 1 | \mathbf{x}) = P(C \leq 1 | \mathbf{x})$$

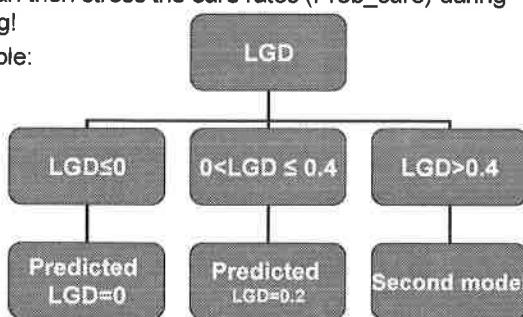
$$P(C = i | \mathbf{x}) = P(C \leq i | \mathbf{x}) - P(C \leq i-1 | \mathbf{x})$$

$$P(C = C_{\max} | \mathbf{x}) = 1 - P(C \leq C_{\max} - 1 | \mathbf{x})$$

63

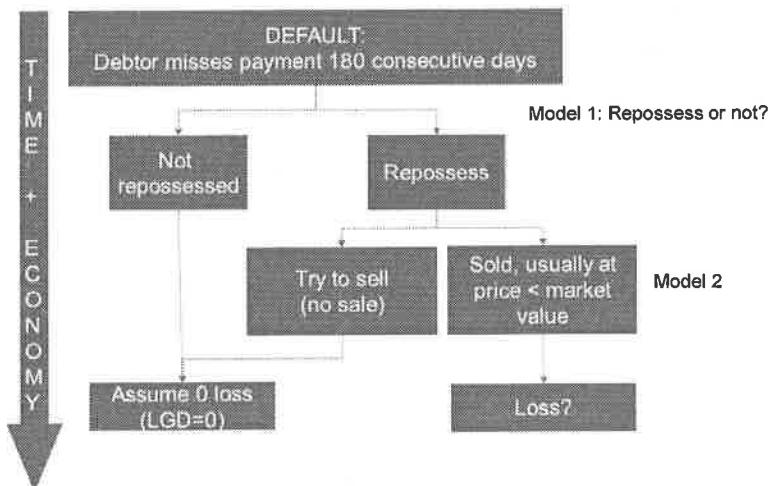
Level 1: Two-Stage Models

- $LGD = Prob_cure \times Fixed + (1 - Prob_cure) \cdot LGD_{workout}$
- Prob_cure is the result of a logistic regression or decision tree model.
- Fixed is the fixed handling cost.
- $LGD_{workout}$ is the result of another predictive model (aka Loss Given Loss).
- Note: You can then stress the cure rates (Prob_cure) during stress testing!
- Other example:



64

Level 1: Two-Stage Model for Mortgages



Leow, Mues, Credit Risk Models for Mortgage Loans, 2011

Level 1: Advanced Models

- Neural networks
- Support Vector Machines
- However, loss of interpretability
- Cf. infra

56

Level 1: Summary of Look-up and Regression Approaches

Sophistication Level	Type	Details	Plus	Minus
Low	Contingency or Look-Up Table	A cell might be: LGD for Sr. unsecured loans for the automotive industry during a recession	Easy to build and use	Very data intensive to completely fill a possibly very large table
Medium	Basic regression	LGD , regressed on dummies for $SP/3$, collateral quality if any (say 3 buckets), industry group (say 6-12), expansion/recession	Relatively easy to build, flexible on data quantity, could easily be converted into a "scorecard"	Grouping/bucketing must be done with care
Medium-high	Advanced regression	As above, but with separate regression models, as warranted, for place in capital structure, collateral quality, expansion/recession; allow for different functional forms (e.g., non-linearity)	Better fit to data	Requires more sophisticated modeling knowledge; somewhat prone to overfitting and datamining
High	Neural nets, tree methods, machine learning	Variety of methods which are often better suited for categorical variables (e.g., place in capital structure, industry) than ordinary regression	Even better fit to data	Even more sophistication; prone overfitting and datamining

57

Level 1: Performance Measures for LGD

y_i – actual LGD; \hat{y}_i – estimated LGD; \bar{y} – average LGD

- R-squared $= 1 - \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$

- Industry benchmarks of R-squared are around 20%–30%!

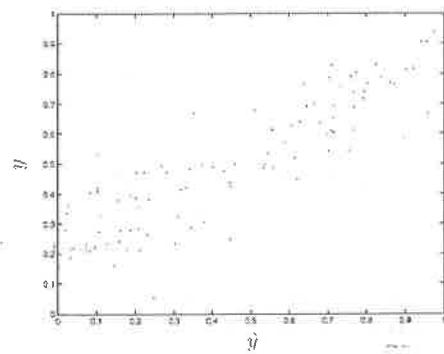
- Mean Squared Error $MSE = \frac{\sum_{i=1}^n (y_i - \hat{y}_i)^2}{n}$

- Mean Absolute Deviation $MAD = \frac{\sum_{i=1}^n |y_i - \hat{y}_i|}{n}$

58

Level 1: Performance Measures for LGD

Scatter plot

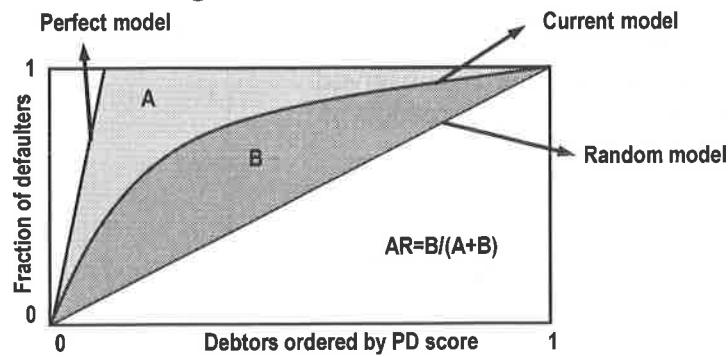


Correlation

$$\text{corr}(\hat{y}, y) = \frac{\frac{1}{n-1} \sum_{i=1}^n (\frac{\hat{y}_i - \bar{\hat{y}}}{s_{\hat{y}}})(\frac{y_i - \bar{y}}{s_y})}{\sqrt{\sum_{i=1}^n (\hat{y}_i - \bar{\hat{y}})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}},$$

59

Level 1: Cumulative Accuracy Profile (CAP) for PD Modeling



Also called Power Curve or
Lorenz Curve!

60

Level 1: Performance Measures for Estimated LGDs

- CAP plot and accuracy ratio 1
 - Binary outcome represents whether observed LGD is higher than the long-term average LGD
 - Indicates how much better model predicts than a long-term average
- CAP plot and accuracy ratio 2
 - Binary outcome represents whether the observed LGD is higher than the long-term 75 percentile of the LGD distribution
 - Indicates how much better the model allows to predict high LGDs
 - Illustrates the performance on the right tail of the LGD distribution

61

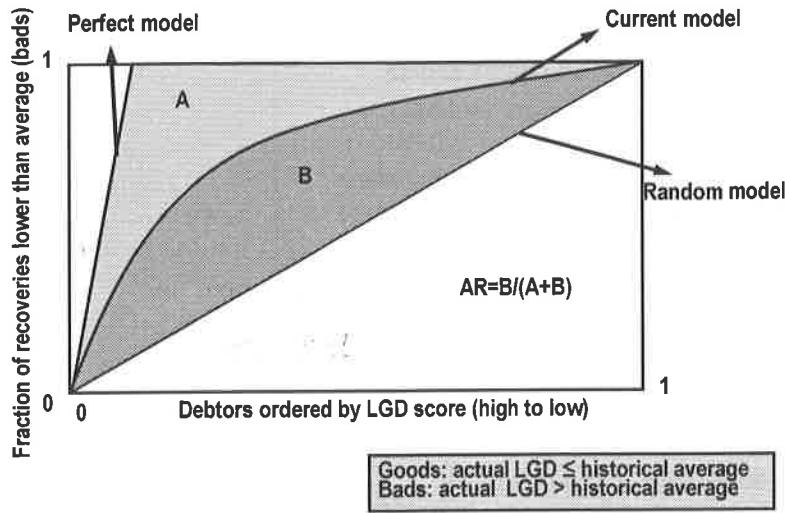
continued...

Level 1: Performance Measures for Estimated LGDs

- CAP plot and accuracy ratio 3
 - Binary outcome represents whether the observed LGD is higher than the long-term 25 percentile of the LGD distribution
 - Indicates how much better the model allows to predict low LGDs
 - Illustrates the performance on the left tail of the LGD distribution

62

Level 1: CAP Curve for LGD Modeling



63

Level 1: Efficiency Measures of Estimated LGDs

- Compute confidence intervals for estimated LGDs
- Width of a confidence interval provides information about the precision and efficiency of the estimate
- Narrower confidence intervals are preferred!
 - More certainty regarding required capital to protect against losses
- Reliability can then be measured as the number of times the actual losses fall outside the confidence interval on an independent test set

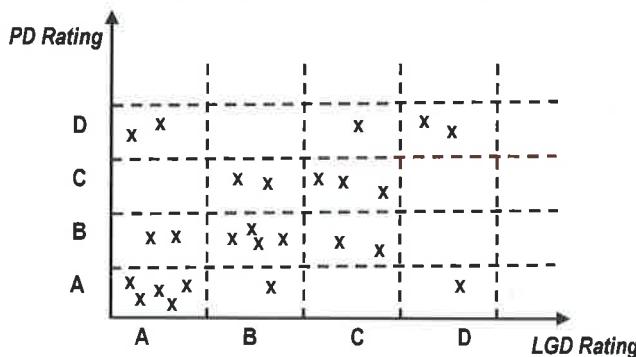
64

Level 2: Mapping to LGD Rating Grades

- Map the output of the regression/segmentation model to LGD facility rating grades
- Compute LGD per rating grade based on historical data
- Additional standards for corporate, sovereign, and bank exposures
 - “Estimates of LGD must be based on a minimum data observation period that should ideally cover at least one complete economic cycle but must in any case be no shorter than a period of **seven years** for at least one source.” (par. 472)
 - “There is **no specific number of facility grades** for banks using the advanced approach for estimating LGD.” (par. 407)
- Additional standards for retail exposures
 - The minimum data observation period for LGD estimates for retail exposures is **five years**. (par 473)

66

Level 2: Mapping to LGD Rating Grades



Check correlations between PD and LGD rating.
Check migration between PD and LGD ratings over time.

66

Level 2: Mapping to LGD Rating Grades

- Create ratings based on LGD score
- Maximize the following formula for K ratings (clusters) (Van Gestel, 2007):

$$\sum_{i=1}^{K-1} \frac{|\overline{LGD}_i - \overline{LGD}_{i+1}|}{\sqrt{\sigma_i^2 + \sigma_{i+1}^2}}$$

- Use external rating scale as a reference

67

Level 2: LGD Ratings and Rating Agencies

Might also be useful for benchmarking.

Moody's	S & P	Fitch
LGD1 91%-100%	1+	100% RR1 91%-100%
LGD2 71%-90%	1	100% RR2 71%-90%
LGD3 51%-70%	2	80%-100% RR3 51%-70%
LGD4 31%-50%	3	50%-80% RR4 31%-50%
LGD5 11%-30%	4	25%-50% RR5 11%-30%
LGD6 0%-10%	5	0-25% RR6 0%-10%

68

Level 2: Paragraph 468 of the Basel II Accord

Standards for all asset classes (advanced approach)

- “A bank must estimate an LGD for each facility that aims to reflect **economic downturn** conditions where necessary to capture the relevant risks. This LGD **cannot be less** than the **long-run default-weighted average loss rate given default** calculated based on the average economic loss of all observed defaults within the data source for that type of facility... a bank must take into account the potential for the LGD of the facility to be higher than the default-weighted average during a period when credit losses are substantially higher than average...this **cyclical variability** in loss severities may be important and banks will need to incorporate it into their LGD estimates... banks may use averages of loss severities observed during periods of high credit losses, forecasts based on appropriately conservative assumptions, or other similar method...using either internal or external data.”

69

Level 2: How to Measure Long-Run LGD

- Historic long-run data
 - Time weighted LGD
 - First calculate LGD for individual years and then average
 - Default weighted LGD
 - Calculated by dividing total losses by the total amount of assets in default
 - Exposure weighted average
 - Weight each default by EAD
 - Default count average
 - Each default has equal weighting

70

Level 2: Averaging LGD

	Default count averaging	Exposure-weighted averaging
Default-weighted averaging	<p>Option 1: Each default has equal weighting defaults from all years grouped into a single cohort</p> $LGD = \frac{\sum_{y=1}^m \sum_{i=1}^{n_y} LR_{i,y}}{\sum_{y=1}^m n_y}$	<p>Option 2: Weighting of each default is determined by exposure at default defaults from all years grouped into a single cohort</p> $LGD = \frac{\sum_{y=1}^m \sum_{i=1}^{n_y} EAD_{i,y} \cdot LR_{i,y}}{\sum_{y=1}^m \sum_{i=1}^{n_y} EAD_{i,y}}$
Time-weighted averaging	<p>Option 3: Each default has equal weighting within annual cohort average average calculated as average of annual averages</p> $LGD = \left[\frac{\sum_{y=1}^m \sum_{i=1}^{n_y} LR_{i,y}}{n_y} \right] \frac{n_y}{m}$	<p>Option 4: Weighting of each default within annual cohort average is determined by exposure at default, average calculated as average of annual averages</p> $LGD = \left[\frac{\sum_{y=1}^m \sum_{i=1}^{n_y} EAD_{i,y} \cdot LR_{i,y}}{\sum_{y=1}^m \sum_{i=1}^{n_y} EAD_{i,y}} \right] \frac{n_y}{m}$

Where,

- i refers to the default observation and y refers to the year/cohort (there are n_y defaults in each year y , and a total of m years of observations)
- EAD is the exposure at default
- LR is the loss rate (Loss amount / EAD) for each observation

71

Level 2: Averaging LGD

- “In most cases, it will not be acceptable to calculate ELGD as the average of annual loss rates.”
(Federal Register)
- “A default-weighted average LGD should not be weighted by exposure size. However, historical data that has been collected on this basis may be acceptable for a transitional period, if firms can demonstrate this does not produce significant differences from averages on a ‘default count’ basis.”
(FSA, CP 06/03: Capital Standards 2, February 2006)

72

Level 2: Example of LGD Averages

Year 1: 20 defaults of \$40 with average loss of 10%

Year 2: 50 defaults of \$100 with average loss of 90%
30 defaults of \$140 with average loss of 60%

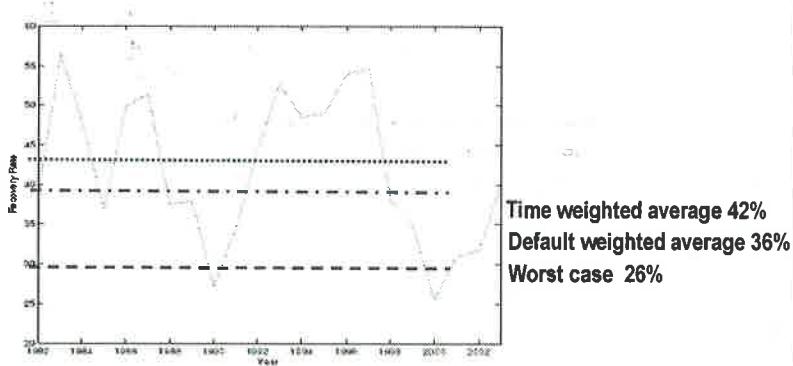
Long run LGD	Default count averaging	Exposure weighted averaging
Default weighted averaging	$\frac{20.10 + 50.90 + 30.60}{20 + 50 + 30} = 67$	$\frac{20.10 \cdot 40 + 50.90 \cdot 100 + 30.60 \cdot 140}{20 \cdot 40 + 50 \cdot 100 + 30 \cdot 140} = 71$
Time weighted averaging	$\frac{20.10 + \frac{50.90 + 30.60}{2}}{2} = 44$	$\frac{50.10 \cdot 90 + 30.14 \cdot 60}{50 \cdot 100 + 30 \cdot 140} + \frac{20.40 \cdot 10}{20 \cdot 40} = 43$

73

→ negotiations won't let doing

this for LGD

Level 2: LGD Drivers: Impact of Economy



: Value Weighted Recovery Rates for Corporate Loans (1982-2003), Moody's KMV

74

Level 2: Guidance on Paragraph 468 of the Framework Document

- BIS LGD working group
- Problems
 - Potential for realized recovery rates to be lower than average during times of high default rates might be a material source of unexpected credit loss (risk to underestimate capital)
 - Data limitations to estimate LGD (and specify economic downturn conditions)
 - Little consensus in industry on how to incorporate downturn conditions into estimation of LGD
- A principles-based approach is suggested

76

continued...

Level 2: Guidance on Paragraph 468 of the Framework Document

Principle 1: Bank must have a rigorous and well-documented process for assessing effects, if any, of economic downturn conditions on recovery rates and for producing LGD estimates consistent with downturn conditions

- Identification of downturn conditions (e.g., negative GDP growth, elevated unemployment rates, ...)
- Identification of adverse dependencies, if any, between default rates and recovery rates
- Incorporation of adverse dependencies, if any, into LGD estimates
 - If adverse dependent: LGD based on averages of loss rates during downturn periods
 - If not adverse dependent: LGD based on long-run default-weighted averages of observed loss rates (neutral conditions)

76

continued...

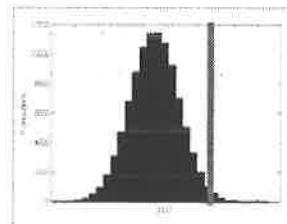
Level 2: Guidance on Paragraph 468 of the Framework Document

- Principle 2: For the estimation of LGDs, measures of recovery rates should reflect the costs of holding defaulted assets over the workout period, including an appropriate risk premium.
- For example, discount the stream of recoveries and workout costs by a risk-adjusted discount rate that is the sum of the risk-free rate and a spread appropriate for the risk of the recovery and cost cash flows.
- Still much confusion/debate about paragraph 468!
 - No clear “sound practice”
 - Ongoing communication between regulators and banks

77

Level 2: Economic Downturn LGD

- Because LGD not based on VAR approximation
- Take average of the x-worst years
- Use upper percentile of the distribution/regression model
- Use mapping formula (e.g., in the U.S.)
 - $LGD_{ed} = 0.08 + 0.92 LGD_{av}$
- Specify downturn scenario and quantify impact on LGD; e.g., look at downturns in regional sub-portfolios and extrapolate
- Bootstrapping (Van Gestel 2008)
 - Create a bootstrap (=sample with replacement from the original data)
 - Compute the mean
 - Repeat this, e.g., 100.000 times
 - This will give the distribution of the mean LGD
 - Management chooses percentile of the distribution



78

Level 2: Economic Downturn LGD (Van Gestel 2008)

If no difference across the periods, then no dependence and no downturn calibration needed.

If there is a difference, use the first row(s)!

	Segment 1	Segment 2	Segment K	
Period 1: (year with highest DR)	avg LGD (period 1, segment 1)	avg LGD (period 1, segment 2)		avg LGD (period 1, segment K)
Period 2: (2 years with highest DR)	avg LGD (period 2, segment 1)	avg LGD (period 2, segment 2)		avg LGD (period 2, segment 3)
Period n-1: (all years except year with lowest DR)	avg LGD (period n-1, segment 1)	avg LGD (period n-1, segment 2)		avg LGD (period n-1, segment K)
Period n: (all years)	avg LGD (period n, segment 1)	avg LGD (period n, segment 2)		avg LGD (period n, segment K)
Reference	calibrated LGD1	calibrated LGD2		calibrated LGD _K

79

↓
Periods with
Decreasing Default Rates

Economic Downturn LGD

- Often not downturn for other business purposes (e.g., economic capital)
 - Violation of use-test but allowed by most supervisors provided the reasons are documented!
- Should it be stress tested (cf. infra)?
 - “When a firm assumes stress and downturn conditions that are similar, the LGD estimates used might also be similar.” (FSA, November 2005)
- “Firms that wish to use estimates of LGD that are zero or close to zero should be able to demonstrate that such a figure is appropriate.” (FSA, November 2005)

80

U.S. Specific

- Originally, the following mapping formula was suggested:
 - $LGD=0.08 + 0.92*ELGD$
- Agencies will not include it in the final rule, but continue to believe that it is a reasonable aid.
- Economic downturn conditions defined as those conditions in which the aggregate default rates for the exposure's entire wholesale or retail subcategory held by the bank in the exposure's national jurisdiction were significantly higher than average.

81

Exposures in Default

- $K=\max(0,LGD-EL_{best})$, whereby LGD is the loss forecast reflecting potential unexpected losses during the recovery period based on downturn conditions, and EL_{best} is the best estimate of expected loss given current economic circumstances and the exposure status
- EL_{best} can be the result of regression model estimating expected loss for exposure given the current status of collection and the economy.
- See paragraph 219 of CP10.

82

Average LGD Estimates for Non-Defaulted Exposures According to QIS5 (June 2006)

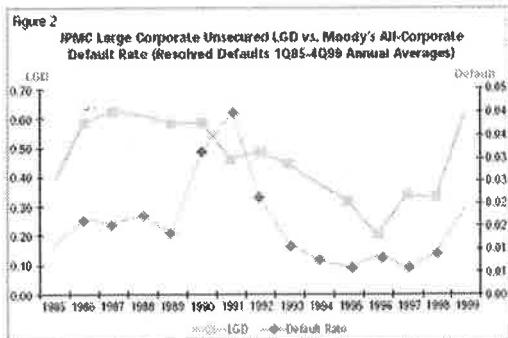
LGD averages for different portfolios in per cent

	IRB Retail				AIRB			
	RM	QRE	Other	SME	Wholesale		SME Corp.	
					Corp.	Bank		
G10 Group 1 (excl US)					39.8	40.9	33.3	35.0
G10 Group 1 (incl US)	20.3	71.6	48.0	46.2				
G10 Group 2	26.2	57.5	43.0	31.1				
CEBS Group 1	16.1	55.0	47.9	38.8	38.1	37.7	27.7	35.1
CEBS Group 2	21.4	51.9	42.2	31.7	35.2	39.4	38.2	26.7
Other non-G10 Group 1	11.0	67.2	48.3	28.4				
Other non-G10 Group 2	40.4	55.7	45.1	49.6				

This table includes banks which participated in QIS 5, as well as additional data for the US. The figures take account of the 10% LGD floor applicable for exposures in the retail residential mortgage portfolio and include only non-defaulted exposure.

83

PD/LGD Correlation



$$\text{LGD} = 0.35 + 7.18 \times \text{Default Rate}, R^2=0.25$$

$$\text{LGD} = 1.16 + 0.16 \times \ln(\text{Default Rate}), R^2=0.40$$

Virtually no correlation between LGD of secured exposures and business cycle!

84

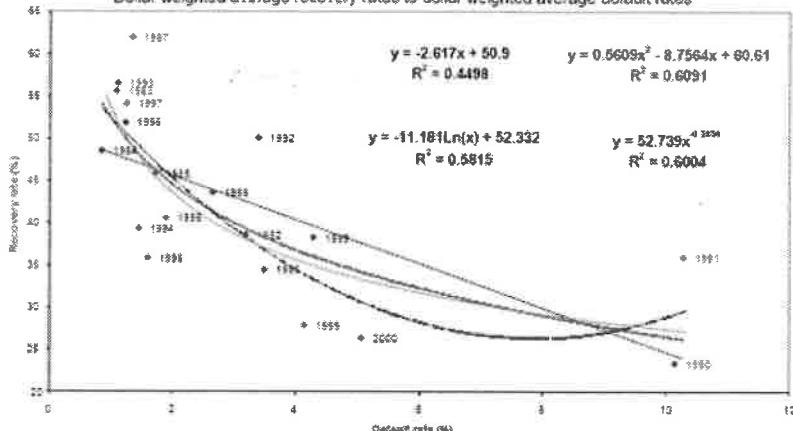
Measuring LGD on Commercial Loans: an 18-year Internal Study,
The RMA Journal, May 2004 (JPMorgan Chase)

PD/LGD Correlation

Recovery rate/default rate association

Altman defaulted bonds data set (1982-2000)

Dollar weighted average recovery rates to dollar weighted average default rates



85

Altman, Brady, Resti, Sironi 2003

Expected Loss versus LGD

- Expected Loss Rate=PD*LGD
- Can back out LGD from EL if PD known!
- Combining PD and LGD

PD/LGD	High	Average	Low
High	7.6%	3%	2.5%
Medium	3.75%	2.5%	1.25%
Low	1.875%	1.25%	0.625%

Accept
Reject

Expected Loss Rate!

86

Exposure at Default (EAD) Modeling

- On-balance-sheet (for example, term loan): nominal outstanding balance, net of specific provisions
- E.g., for installment loans
 - EAD is the amount outstanding at the time of capital calculation.
- No downward adjustment for amortization or expected prepayments
 - EAD is floored at current outstanding.
- Off-balance-sheet (for example, credit cards, revolving credit): committed but unused loan amount times a credit conversion factor (CCF)
- CCF is needed to take into account additional drawings prior to default.
- Also called Loan Equivalency Factor (LEQ) in the U.S.

87

EAD Modeling for Revolving Credit

- $EAD = DRAWN + CCF \times (LIMIT - DRAWN)$
- CCF is the proportion of the undrawn but committed amount likely to be drawn prior to default; also called credit conversion factor (CCF), or loan equivalency factor (LEQ).
- $0 \leq CCF \leq 1$
- Conservative estimates set CCF to 1.
- *"The EG proposes that it should be up to firms to determine and justify their approach for estimation of CCFs and for supervisors to review and agree these approaches."* (FSA, June 2006)

88

EAD According to Basel II

For corporates/sovereigns/banks:

- For off-balance sheet items, exposure is calculated as the committed but undrawn amount multiplied by a CCF. There are two approaches for the estimation of CCFs: a foundation and an advanced approach.” (par. 310).
- Foundation approach:
 - “Commitments with an original maturity up to one year and commitments with an original maturity over one year will receive a CCF of 20% and 50%, respectively. However, any commitments that are unconditionally cancelable at any time by the bank without prior notice,, will receive a 0% CCF.” (par. 83)

89

EAD According to Basel II

For retail:

- “Both on- and off-balance sheet retail exposures are measured gross of specific provisions or partial write-offs. The EAD on drawn amounts should not be less than the sum of (i) the amount by which a bank’s regulatory capital would be reduced if the exposure were written-off fully, and (ii) any specific provisions and partial write-offs.” (par. 334)
- “For retail off-balance sheet items, banks must use their own estimates of CCFs.” (par. 335)

90

continued...

EAD According to Basel II

- "For retail exposures with uncertain future drawdown such as credit cards, banks must take into account their history and/or expectation of additional drawings prior to default in their overall calibration of loss estimates. In particular, where a bank does not reflect conversion factors for undrawn lines in its EAD estimates, it must reflect in its LGD estimates the likelihood of additional drawings prior to default. Conversely, if the bank does not incorporate the possibility of additional drawings in its LGD estimates, it must do so in its EAD estimates." (par. 336)
- "Advanced approach banks must assign an estimate of EAD for each facility." (par. 475)
- Must use margin of conservatism and economic downturn EAD if EAD volatile over economic cycle. (Federal Register)
- The minimum data observation period for EAD estimates for retail exposures is five years (seven years for corporate exposures). (par. 478 and 479)

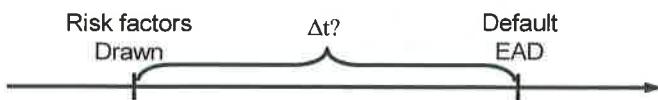
91

Definition of EAD

- **Definition 1**
 - Total exposure at risk on the very moment of default
 - Additional drawings considered as a cost and enter in the LGD (might cause LGD to be > 100%)
 - EAD fixed at time of default, LGD dependent on length of recovery process
- **Definition 2**
 - Maximum observed EAD during the default process
 - Definition also takes into account drawings after default (for example, given by bank with the perspective of a cure or reduced loss)
 - LGD almost sure < 100% (except with additional unrecovered costs)
 - Both EAD and LGD dependent on length of recovery process
 - Problem is which time point to choose for discounting
- Additional extensions of credit after default should be included in LGD. (Federal Register)

92

EAD Modeling



- Need two time points:
 - EAD measured at time of default
 - Drawn balance and risk factors measured at Δt before default
 - Can then back out CCF as follows:

$$CCF = \frac{EAD - drawn}{Limit - drawn}$$

- Problem is how to determine Δt

93

EAD Modeling

Cohort method

- Group defaulted facilities into discrete calendar periods (for example, 12 months unless other time period more conservative and appropriate) according to date of default
- Collect information about risk factors and drawn/undrawn amount at beginning of calendar period and drawn amount at date of default
- Pool data of different calendar periods for estimation
- For example, calendar period is defined as 1 November 2009 to 31 October 2010, and then information about risk factors and drawn/undrawn amount on 1 November 2009 should be collected and drawn amounts of facilities upon default.

94

EAD Modeling

Fixed-horizon method

- Collect information about risk factors and drawn/undrawn amount for a fixed time interval prior to the date of default (at least 12 months unless other time period more conservative and appropriate) and the drawn amount on date of default, regardless of the actual calendar date on which the default occurred.
- For example, fixed interval is 12 months; if a default occurred on 15 July 2010, then information about risk factors and drawn/undrawn amount of the defaulted facilities on 15 July 2009 is used.

95

EAD Modeling

Variable time horizon approach

- Variant of fixed time horizon approach using several reference times within the chosen time horizon
- For example, compare drawn amount at time of default with risk factors and drawn/undrawn amounts one month, two months, three months before default

Momentum method

- Express CCF as a percentage of the total limit at time of default
- Drawn amount at time of default is compared to total limit at time of default
- Not allowed because currently drawn amount is not considered (CEBS, CP10, 2006)
- If allowed, $EAD = \max(\text{Drawn}, \text{CCF} \cdot \text{Limit})$

96 What if credit limit changes?

Limits for CCF

- Negative CCF
 - CCF can be negative when borrower has paid back portion of the amount prior to default
 - Truncate in the data for estimation
 - Estimated CCFs cannot be negative (Federal Register)
- CCF > 1
 - Reason: e.g., off-line transactions
 - Problem: as drawn balance increases, exposure decreases
 - Suppose limit=2500, CCF=110%
 - Drawn=1000 Euros, → EAD=1000+1.10*1500=2650 Euro
 - Drawn=1500 Euros → EAD=1500+1.10*1000=2600 Euro
 - Solution: Convert soft credit limit into hard credit limit based on historical data such that CCF always ≤ 1 (use, for example, 99% confidence level if needed)

97

Risk Factors for Predictive Modeling of EAD (CCF)

Type of obligor

- Corporates and banks: credit lines often not completely utilized at time of default
- Retail and SMEs: more likely to overdraw (or fully utilize) credit line

A borrower's access to other sources of funding

- Retail and SME have fewer access to alternative sources than large corporate obligors and banks (use type of obligor as proxy)

Factors affecting the borrower's demand for funding/facilities

Expected growth in a borrower's business (and accordingly rise in funding requirements)

The nature of the particular facility (for example, industry, geographical region, facility size, covenant protection, credit risk (e.g., PD, rating, behavioral score), ...)

98

Developing EAD Models

- Level 0/Level 1/Level 2 (see LGD modeling)
- Level 0
 - Construct development data set storing information (risk factors + CCF) of the defaulted facilities
 - Use same definition of default as for PD and LGD
- Level 1
 - Develop segmentation/regression/two-stage model
 - U-shaped distribution observed before; use beta-distribution and same trick as with LGD
 - Same performance measures as for LGD (R-squared, MSE, CAP plots, ...)
- Level 2
 - For defining ratings and economic downturn CCF:
See LGD section

Case Study: CCF Modeling for U.K. Bank

Brown I., Mues C., Regression Model Development for Credit Card Exposure at Default, submitted, 2011.

Data set obtained from a major U.K. financial institution and contains monthly data on credit card usage for a three-year period (January 2001 – December 2004).

Bi-modal CCF distribution was observed



100

Case Study: Variables Used for CCF Model

Moral (2006) Variables

- **Committed amount** – the advised credit limit at the start of cohort
- **Drawn amount** – the exposure at the start of cohort
- **Undrawn amount** – the advised limit minus the exposure at the start of cohort
- **Credit percentage usage** – the exposure at the start of cohort divided by the advised credit limit at the start of cohort
- **Time to default** – number of months between start of cohort and default date
- **Rating class** – the behavioral score at the start of the cohort, binned into four categories

101

continued...

Case Study: Variables Used for CCF Model

Additional Variables

- **Average number of days delinquent** – Previous 3, 6, 9, and 12 months prior to the start of the cohort
- **Increase in committed amount** – Binary variable indicating whether there has been an increase in the committed amount since 12 months prior to the start of the cohort
- **Undrawn percentage** – The undrawn amount at the start of the cohort divided by the advised credit limit at the start of the cohort
- **Relative change in drawn, undrawn and committed amount**
- **Absolute change in drawn, undrawn and committed amount**

Moral G., EAD Estimates for Facilities with Explicit Limits. In: Engelmann B, Rauhmeier R (Eds), *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, Springer, Berlin, 197–242, 2006.

102

Case Study: Results of Regression Approaches

Variables	OLS model (using only Alford (2006) suggested variables)		OLS model (additional variables)		Binary logit model (LOGIT)		Cumulative logit model (CLOGIT)	
	Parameter Estimate	P-value	Parameter Estimate	P-value	Parameter Estimate	P-value	Parameter Estimate	P-value
Intercept 1							0.3493	<0.001
Intercept 2							-0.5491	<0.01
Credit percentage usage	-0.1229	<0.001	-0.1269	<0.001	-0.5737	<0.001	-1.322	<0.001
Committed amount	1.73E-05	<0.0001	1.76E-05	<0.0001	9.0E-05	<0.001	8.8E-05	<0.001
Drawn								
Undrawn	-8.68E-05	<0.001	-8.88E-05	<0.001	-4.7E-04	<0.001	-3.6E-04	<0.001
Time-to-Default	0.0334	<0.001	0.0316	<0.001	0.1538	<0.001	0.1069	<0.001
Rating class								
Rating 1 (AAA-A)	0.1735	<0.001	0.3364	<0.001	0.4000	0.0069	-0.0772	0.5471
Rating 2 (BBB-B)	0.3483	<0.001	0.5977	<0.001	0.5885	<0.001	0.6922	<0.001
Rating 3 (C)	0.5944	<0.001	0.1201	<0.001	-0.2121	0.0043	-0.0157	0.8098
Rating 4 (UR)								
Average number of days delinquent in the last 6 months			0.0048	<0.001	0.0216	<0.001	0.0218	<0.001
Coefficient of Determination (R^2)	0.0981		0.0960		0.1628		0.0833	
Fearnley's Correlation Coefficient (λ')	0.3170		0.3144		0.3244		0.3897	
Fearnley's Correlation Coefficient (ρ')	0.3932		0.2943		0.3283		0.2943	
Root Mean Squared Error (RMSE)	0.4393		0.4398		0.4704		0.4432	

103

Correlation between PD/LGD/EAD

- Correlation of PDs across borrowers
 - Related to asset correlation parameter in Basel II Accord
 - Determined using some empirical but not published procedure
- Positive correlation between PD and LGD
 - PDs usually higher during economic downturns when asset values get depressed also, hence higher than average LGDs
 - As correlation increases, so will level of credit risk
 - Treated as independent in Basel II
- Correlation between PD and EAD
 - For example, revolving credits, if financial distress worsens, a borrower will draw down as much as possible on existing unutilized facilities in order to avoid default
 - When EAD depends on market prices (for example, traded bonds)
- Impact on, for example, stress testing

104

Relations between Basel Parameters

	Some consensus	Open questions
Probability of default (PD)	+ correlation with asset values. Time-varying with systematic risk component.	Relationship between PD correlations and firm credit quality (PD level). Relationship between PD correlations and bankruptcy rates (renegotiation) and macroeconomic shifts.
Loss given default (LGD)	+ correlation with asset collateral values. Time-varying with systematic risk component.	Relationship between LGD correlations and firm credit quality (PD level). Relationship between LGD correlations and bankruptcy rates (renegotiation) and macroeconomic shifts.
Correlations between LGD and PD		Sign of correlation between LGD and PD. Relationship between PD-LGD correlation and systematic macro effects.
Exposure at default (EAD)	Time-varying with systematic risk component. -	Sign of interim correlation in EAD. Relationship between EAD correlations and firm credit quality (PD level). Relationship between EAD correlations and bankruptcy rates (renegotiation) and macroeconomic shifts. Integration of market risk and credit risk models.

BIS Working Paper 126 2003

Chapter 14 Validation of Basel II Models

14.1 Validating Basel II Models.....14-3

14.1 Validating Basel II Models

Validation According to Basel II

- “The bank must have a **regular cycle of model validation** that includes monitoring of model performance and stability; review of model relationships; and testing of model outputs against outcomes.” (par. 417)
- “Banks must have a robust system in place to validate the **accuracy and consistency of rating systems, processes** and the estimation of all relevant risk components.” (par. 500)
- “Banks must regularly **compare realized default rates** with **estimated PDs** for each grade and be able to demonstrate that the realized default rates are within the expected range for that grade... (similar for LGD and EAD) ... This analysis and documentation must be updated at least **annually**.” (par. 501)

3

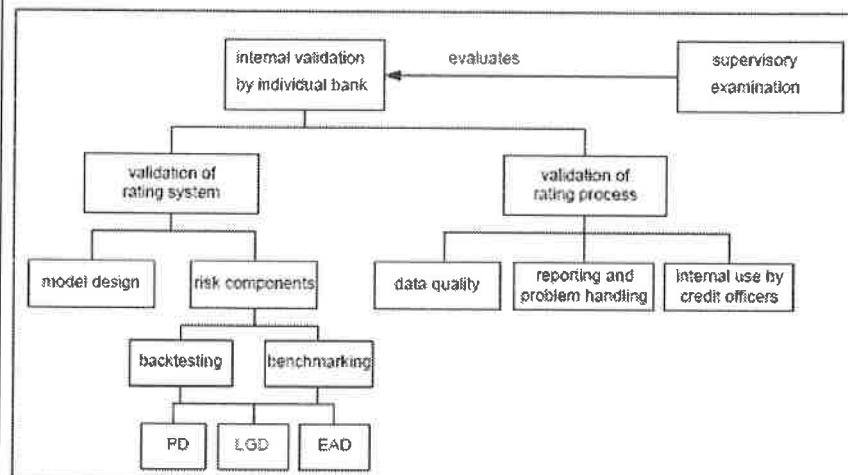
continued...

Validation According to Basel II

- “Banks must also use other **quantitative validation tools** and comparisons with relevant **external data sources**.” (par. 502)
- “Banks must have **well-articulated internal standards** for situations where **deviations** in realized PDs, LGDs and EADs from expectations become significant enough to call the validity of the estimates into question.” (par. 504)

4

Validation According to Basel II



BIS 14, Working paper

6

Validation According to the Federal Register

- “Banks must have an effective system of controls that ensures ongoing compliance with the qualification requirements, maintains the integrity, reliability and accuracy of the IRB system, and includes adequate corporate governance and project management processes.”
- “Validation must assess the **accuracy** of the risk rating and segmentation systems and the quantification process.”
- “Validation processes for risk rating and segmentation systems, and the quantification process must include the **evaluation of conceptual soundness, ongoing monitoring, and outcomes analysis**.”

6

continued...

Validation According to the Federal Register

- “Banks must **benchmark** their risk rating and segmentation systems, and their risk parameter estimates.”
- “Banks must analyze outcomes and must develop statistical methods to **backtest** their risk rating and segmentation systems and the quantification process.”
- “Banks should establish ranges around the estimated values of risk parameter estimates and model results in which actual outcomes are **expected to fall** and have a validation policy that requires them to assess the reasons for differences and that outlines the timing and type of **remedial actions** taken when results fall outside expected ranges.”

7

Validation Terminology: Backtesting versus Benchmarking

- Backtesting
 - Using statistical (quantitative) methods to compare estimates of PD, LGD, and EAD to realized outcomes

Rating Category	Estimated PD	Nr. of observations	Nr. of observed defaults
A	2%	1000	17
B	3%	500	20
C	7%	400	35
D	20%	200	50

- Benchmarking
 - Comparing internal estimates across banks and/or with external benchmarks
- Validation > Backtesting + Benchmarking!

8

Quantitative versus Qualitative Validation

- Quantitative validation
 - Backtesting
 - Benchmarking
 - Data + Statistics!
- Qualitative validation
 - Data quality
 - Use test
 - Model design
 - Documentation
 - Corporate governance and management oversight

9

Common Validation Issues

- Banks employ a wide range of techniques to validate internal ratings. The techniques used to assess corporate and retail ratings are substantially different.
- Ratings validation is not an exact science. Absolute performance measures are considered counterproductive by some institutions.
- Expert judgment is critical. Data scarcity makes it almost impossible to develop statistically based internal-ratings models in some asset classes.
- Data issues center around both quantity and quality. Default data, in particular, is insufficient to produce robust statistical estimates for some asset classes.

10

General Validation Principles According to the Basel Committee Validation Subgroup

- Principle 1:** Validation is fundamentally about assessing the predictive ability of a bank's risk estimates and the use of ratings in credit processes.
- Principle 2:** The bank has the primary responsibility for validation.
- Principle 3:** Validation is an iterative process.
- Principle 4:** There is no single validation method.
- Principle 5:** Validation should encompass both quantitative and qualitative elements.
- Principle 6:** Validation processes and outcomes should be subject to independent review.

11

Additional Validation Principles

- Supervisor reviews the bank's validation processes
- Make validation constructive, not threatening
- Independent staff included in the validation process
- Validation does not provide a fixed decision but rather a suggestion for further action and study
- Develop validation frameworks + action schemes
- Validation methods not allowed to change with economic cycle unless clearly and thoroughly documented. (par. 503, Basel II Accord)

12

Developing a Validation Framework

- Diagnose validation needs
 - Work out validation activities
 - Timetable for validation activities
 - Tests and analyses to be performed
 - Actions to be taken in response to findings
 - Why/What/Who/How/When
 - Should be described in a validation policy!

13

Validation Scorecard Presented in FSA CP189

Use test scorecards Data Accuracy Scorecards

"As a result of this consultation we have decided not to implement the use test 'scorecard' as we think it places an unnecessary burden on firms given the range of models being implemented and that we can achieve the same aim by other means." (FSA CP 05/03, par. 7.24)

"At present, we are not planning to proceed with a formal validation scorecard as set out in CP189." (FSA CP 05/03, par. 7.53)

14

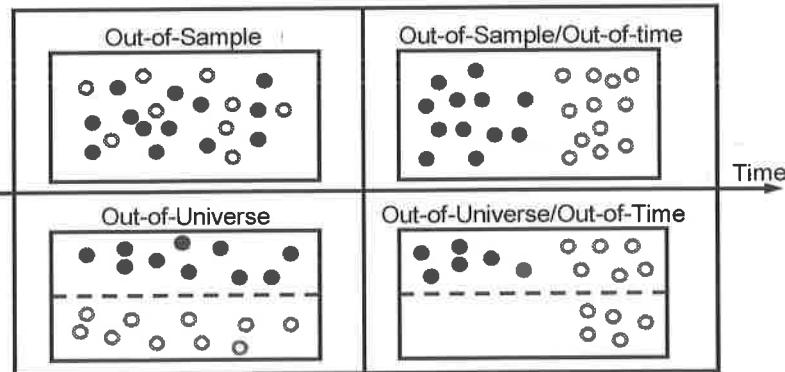
Quantitative Validation

- Measure how well ratings predict default risk, loss, or exposure
 - PD, LGD, CCF
 - “The burden is on the bank to satisfy its supervisor that a model or procedure has good predictive power and that regulatory capital requirements will not be distorted as a result of its use.” (par. 417)
- Compare realized numbers to predicted numbers
- Use appropriate performance metrics and test statistics
- Decide on significance levels!
- Out-of-sample versus out-of-time

16

Out-of-Sample versus Out-of-Time versus Out-of-Universe Quantitative Validation

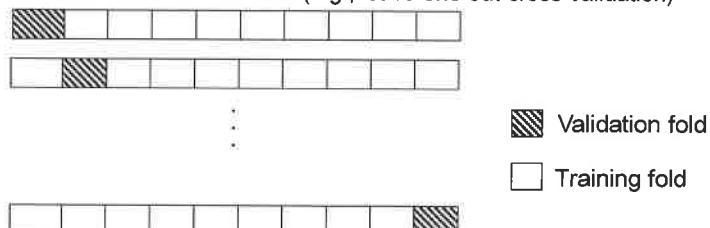
● : Training set ○ : Test set



16

Validation During Model Development versus Validation During Model Usage

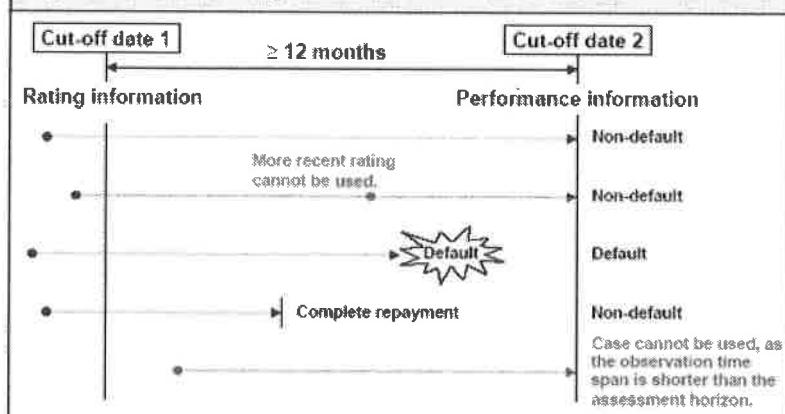
- Validation during model development
 - Typically Out-of-sample
- Validation during model usage
 - Automatically Out-of-sample/Out-of-time (sometimes also Out-of-universe)
- Note
 - Validation during model development: for small samples, use cross-validation methods (e.g., leave-one-out cross-validation)



17

Generating the Data Set for Validation

Figure A1. Generating the data set for validation



18

HKMA, 2006

Problems with Quantitative Validation

- Different sources of variation
 - Sample variation
 - External effects (for example, macro-economy)
 - Internal effects (for example, strategy change)
- Low statistical confidence
 - Suppose we only look at sample variation and the PD for a grade is 100bp, and we want to be 95% confident that actual PD is no more than 20bp off from that estimate.

$$n = \left(\frac{1.96 \sqrt{PD(1-PD)}}{0.002} \right)^2$$

- Would need about 9500 obligors in that grade!
- Statistical independence assumption violated
 - Correlation between defaults
 - Correlation between PD /LGD/EAD
- Data availability!

19

Levels of Backtesting

Calibration:

Mapping of rating to a quantitative risk measure. A rating system is considered well-calibrated if the (ex-ante) estimated risk measures deviate only marginally from what has been observed ex-post.

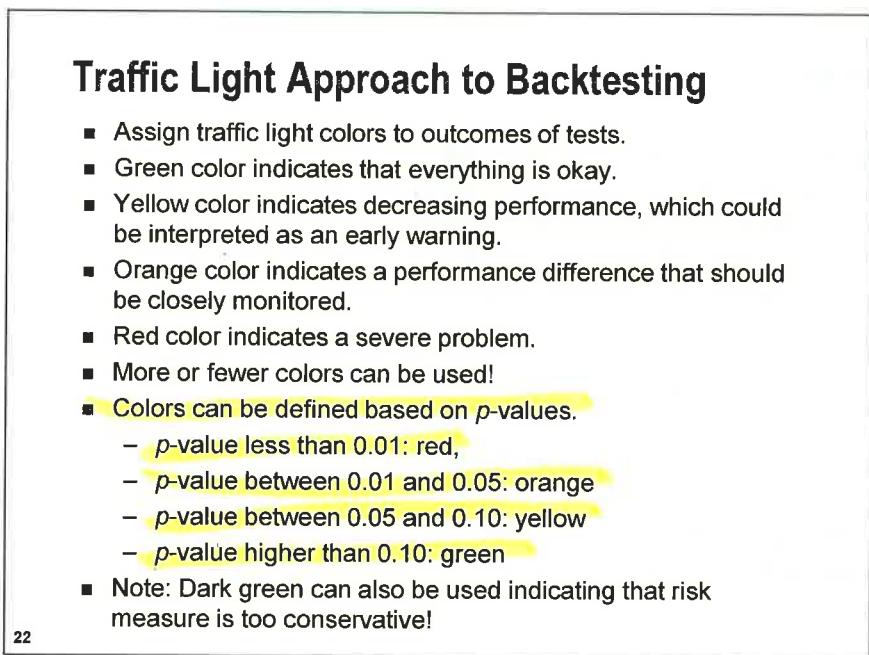
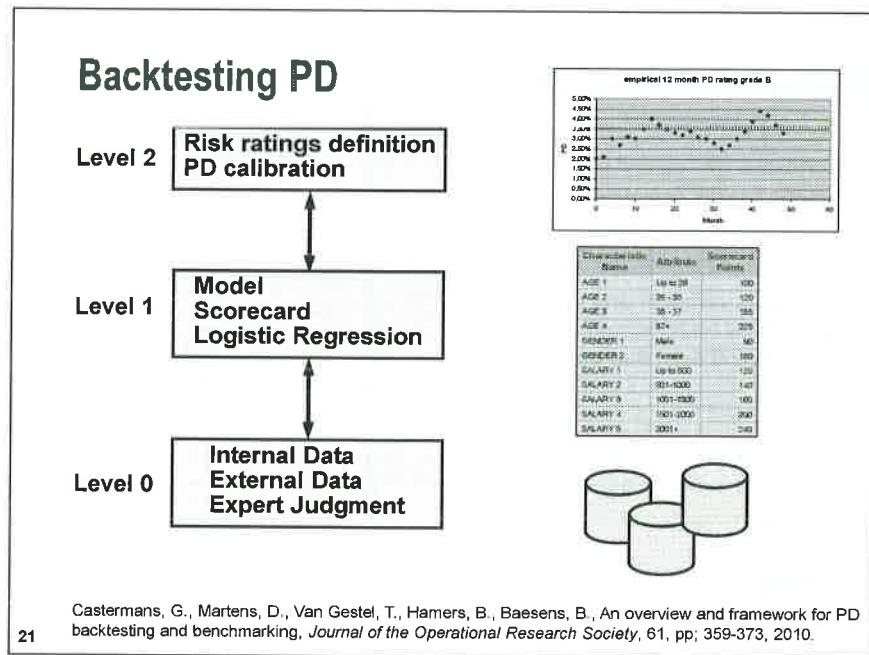
Discrimination:

Measures how well the rating system provides an ordinal ranking of the risk measure considered.

Stability:

Measures to what extent the population that was used to construct the rating system is similar to the population that is currently being observed.

20



Example Traffic Light Implementation

PD	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av
	0.26%	0.17%	0.42%	0.53%	0.54%	1.36%	2.46%	5.76%	8.76%	20.89%	3.05%
DR	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av
1993	0.00%	0.00%	0.00%	0.83%	0.00%	0.76%	3.24%	5.04%	11.29%	28.57%	3.24%
1994	0.00%	0.00%	0.00%	0.00%	0.00%	0.59%	1.55%	3.75%	7.95%	5.13%	1.38%
1995	0.00%	0.00%	0.00%	0.00%	0.00%	1.76%	4.35%	6.42%	4.06%	11.57%	2.51%
1996	0.00%	0.00%	0.00%	0.00%	0.00%	0.90%	1.17%	0.00%	3.28%	13.99%	0.78%
1997	0.00%	0.00%	0.00%	0.00%	0.00%	0.47%	0.00%	1.54%	7.22%	14.87%	1.41%
1998	0.00%	0.31%	0.00%	0.00%	0.62%	1.12%	2.11%	7.54%	5.52%	15.09%	2.85%
1999	0.00%	0.00%	0.34%	0.47%	0.00%	2.00%	3.28%	6.91%	9.83%	20.44%	3.34%
2000	0.28%	0.00%	0.97%	0.94%	0.63%	1.04%	3.24%	4.10%	10.88%	19.65%	3.01%
2001	0.27%	0.27%	0.00%	0.51%	1.00%	2.82%	3.19%	13.04%	16.39%	34.40%	9.49%
2002	0.35%	0.17%	0.35%	0.68%	1.41%	1.58%	2.00%	6.81%	8.86%	21.43%	3.71%
Av	0.26%	0.17%	0.42%	0.53%	0.54%	1.36%	2.46%	5.76%	8.76%	20.9%	3.05%

23

Backtesting PD at Level 0

- Check whether internal or external environmental changes will impact the rating model
 - New developments in economic, political, or legal environment, changes in commercial law, or bankruptcy procedures (external)
 - Change of business strategy, exploration of new market segments, changes in organizational structure (internal)
- Two-step approach
 - Step 1: check whether population on which model is currently being used is similar to population that was used to develop model
 - Step 2: if differences occur in Step 1, verify stability of individual variables

24

Backtesting PD at Level 0: Step 1

- Construct system stability index (SSI, also called deviation index in SAS) across ratings or score ranges

Population Stability Report

Score range	Expected (training)%	Observed (Actual)%	SSI
0-169	6%	7%	0,0015
170-179	10%	8%	0,0045
180-189	9%	7%	0,0050
190-199	12%	9%	0,0086
200-209	12%	11%	0,0009
210-219	8%	11%	0,0096
220-229	7%	10%	0,0107
230-239	8%	12%	0,0162
240-249	12%	11%	0,0009
250+	16%	14%	0,0027
	100%	100%	0,0605

System Stability Index=

$$\sum (A - T) \ln \frac{A}{T}$$

Rule of Thumb

- < 0.10 : No significant shift (Green flag)
- 0.10 – 0.25 : Minor shift (Yellow flag)
- > 0.25 : Significant shift (Red flag)

25

Backtesting PD at Level 0: Step 1

Score range	Expected (training)%	Observed (Actual)% at t	Observed (Actual)% at t+1
0-169	6%	7%	6%
170-179	10%	8%	7%
180-189	9%	7%	10%
190-199	12%	9%	11%
200-209	12%	11%	10%
210-219	8%	11%	9%
220-229	7%	10%	11%
230-239	8%	12%	11%
240-249	12%	11%	10%
250+	16%	14%	15%
SSI versus Expected		0,0605	0,0494
SSI versus t-1			0,0260

26

Backtesting PD at Level 0: Step 2

Use histograms, *t*-tests, or SSI to detect shifts in variables.

	Range	Expected (training)%	Observed (Actual)% at t	Observed (Actual)% at t+1
Income	0-1000	16%	18%	10%
	1001-2000	23%	25%	12%
	2001-3000	22%	20%	20%
	3001-4000	19%	17%	25%
	4001-5000	15%	12%	20%
	5000+	5%	8%	13%
SSI reference		0,029		0,208
SSI t-1				0,238
Years client	unknown client	15%	10%	5%
	0-2 years	20%	25%	15%
	2-5 years	25%	30%	40%
	5-10 years	30%	30%	20%
	10+ years	10%	5%	20%
	SSI reference	0,075		0,201
SSI t-1				0,262

27

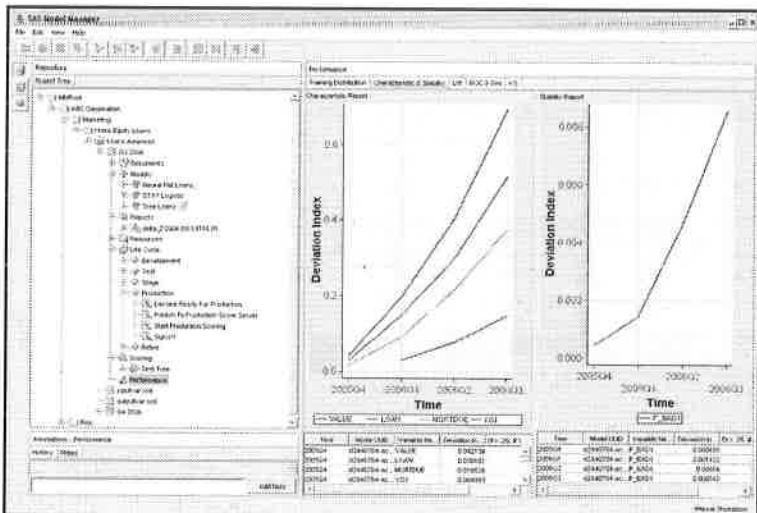
Backtesting PD at Level 0: Step 2

Characteristic analysis

Age	Training	Actual	Points	Index
18-24	12%	21%	10	0,9
25-29	19%	25%	15	0,9
30-37	32%	28%	25	-1
38-45	12%	6%	28	-1,68
46+	25%	20%	35	-1,75
				-2,63

28

Model Manager



29

Backtesting PD at Level 1

- Scorecard level!
- Validation of the logic behind the model used (for example, assumptions)
- For example, for (logistic) regression: qualitative checks on inputs to confirm whether the signs are as expected
- Inspect p-values, model significance, ...
- Input selection (multicollinearity)
- Missing values and outliers
- Coarse classification

30

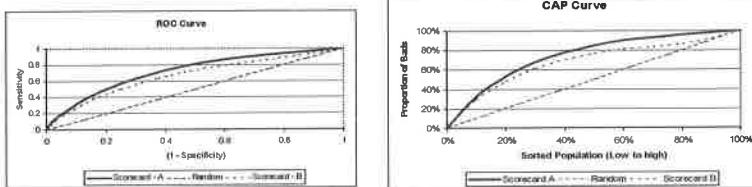
Backtesting at Level 1

- Training set/Test set (hold out set)
 - Out of sample/out of time
- K-fold cross-validation
 - Split data into k-folds (for example, k=10)
 - Estimate algorithm k-times, each time using another fold for calculating the performance
 - Average performance
- Bootstrapping
 - resampling with replacement from the original sample

31

Backtesting at Level 1

- Receiver Operating Curve (ROC) versus Cumulative Accuracy Profile (CAP)



- Area under ROC curve (AUC) versus Accuracy Ratio (AR)
- Accuracy Ratio=Gini coefficient
- $AR=2 \times AUC - 1$
- Confidence intervals around AR/AUC
- "Need at least 50 defaults to successfully calculate AUC and AR" (B. Engelmann, E. Hayden, and D. Tasche 2003)
- "Most of the models we tested had ARs in the range of 50% to 75% for (out-of-sample and out-of-time) validation tests." (Moody's) ...

32

Backtesting at Level 1

AUC	AR	Quality
$0 < \text{AUC} < 0.5$	$\text{AR} < 0$	No discrimination
$0.5 < \text{AUC} < 0.7$	$0 < \text{AR} < 0.4$	Poor discrimination
$0.7 < \text{AUC} < 0.8$	$0.4 < \text{AR} < 0.6$	Acceptable discrimination
$0.8 < \text{AUC} < 0.9$	$0.6 < \text{AR} < 0.8$	Excellent discrimination
$0.9 < \text{AUC} < 1$	$0.8 < \text{AR} < 1$	Exceptional

- ✍ Depends on type of application.
- Application scores have Gini's 0.4-0.6.
- Behavioral scores have Gini's 0.55-0.75.

33

Backtesting PD at Level 1

	AR	Nr of observations	Nr of defaults	Traffic light
AR model				
AR year t				
AR year t+1				
AR year t+2				
...				
Average AR period 1				
Average AR period 2				

34

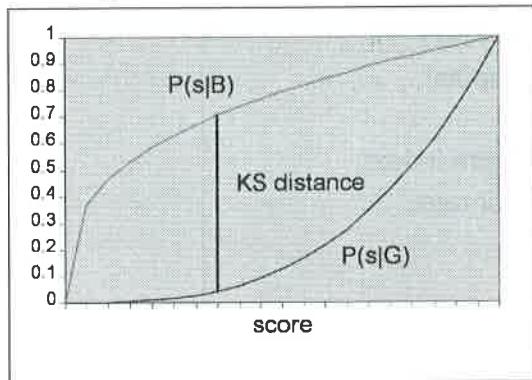
Example Backtesting PD at Level 1

	Number of Obs.	Number of defaulters	AR
AR model	5866	105	0,85
AR 2006	5077	97	0,81
AR 2005	5462	108	0,80
AR 2004	5234	111	0,83
AR 2003	5260	123	0,79
AR 2002	5365	113	0,79
AR 2001	5354	120	0,75
AR 2000	5306	119	0,82
AR 1999	4970	98	0,78
AR 1998	4501	62	0,80
AR 1997	3983	60	0,83
Average AR	5111,2	101,1	0,80

35

Backtesting at Level 1

- Kolmogorov-Smirnov statistic (related to ROC curve)



- Pietra index, Bayesian error rate, conditional entropy, divergence, ...

continued...

36

Backtesting at Level 1

- “The group has found that the **Accuracy Ratio (AR)** and the **ROC measure** appear to be more meaningful than the other above-mentioned indices because of their statistical properties.” (BIS 14 Working Paper)
- “Practical experience shows that the Accuracy Ratio has tendency to take values in the range **50%** and **80%**. However, such observations should be interpreted with care as they seem to strongly depend on the composition of the portfolio and the numbers of defaulters in the sample.” (BIS 14 Working Paper)
- There is **no absolute KS statistic, GINI coefficient, or ROC measure** that models need to reach to be considered adequate. (Monika Mars, Global Internal Ratings Validation Survey, 2003)

37

continued...

Backtesting at Level 1

- It is difficult and probably inappropriate to rely on a single measure, such as the widely used Gini coefficient. (FSA CP 05/03, par. 7.57)
- All depends on predictive power of inputs!
- Nice add-ons to consider:
 - Kolmogorov-Smirnov statistic (Pietra index)
 - Entropy measures, Bayesian Error rate, ... (see BIS14)
 - Scorecard report (overrulings)
 - Lift curves

38

Overrides

- “For model-based ratings, the bank must have guidelines and processes for monitoring cases where human judgement has overridden the model’s rating, variables were excluded, or inputs were altered. Banks must identify overrides and separately track their performance.” (par. 428)
- See also Federal Register.

39

Overrides

Override Report

Score Interval	Accepts	Rejects	Total
<100	<i>1</i>	40	41
100-120	2	30	32
120-140	2	25	27
140-150	3	20	23
150-160	30	2	32
160-170	40	2	42
170-180	30	<i>1</i>	31
Total	108	120	228

- The black line denotes the cutoff; overrides are denoted in italics.
- Number of low-side overrides = $1+2+2+3=8$
- Number of high-side overrides = $2+2+1=5$
- Low-side overrides should be separately monitored

40

Overrides

Binary targets

Original	Override	Actual
A	A	Good
B	B	Good
A	C	Bad
C	C	Bad
A	B	Good
B	B	Good
B	B	Good
B	D	Good
C	C	Bad

- Compare ROCs
- For example, use test of DeLong, DeLong, Clarke-Pearson
- Classification accuracy, notch difference graph
- McNemar test

41

Benchmark to external party

Original	Override	Benchmark
A	A	A
B	B	B
A	C	B
C	C	D
A	B	B
B	B	B
B	B	C
B	D	C
C	C	C

Backtesting at Level 1: Validation Scorecard

COMBINED APPROACH TABLE

Validation Range		Statistical	1.3%	88.33%	43.77%	50.34%	78.17%	1.4%	0.7%
Lower to Upper Limits	Meaning	Mean Difference	CND% > 50% CD: (1-PH)	K-S Statistic	Accuracy Ratio	ROC Statistic	Information Statistic	Kullback-Leibler	
0 1	Random	0.00	50.00%	0.00%	0.00%	50.00%	0.0000	0.0000	
1 2	Doubtful	0.28	59.87%	9.95%	14.08%	57.00%	0.0625	0.0313	
2 3	Poor	0.50	69.15%	19.74%	27.60%	63.80%	0.2500	0.1250	
3 4	Marginal	0.75	77.34%	29.23%	40.40%	70.20%	0.5625	0.2813	
4 5	Satisfactory	1.00	84.13%	38.20%	52.00%	76.00%	1.0000	0.5000	
5 6	Good	1.25	69.44%	46.80%	62.30%	81.15%	1.5625	0.7813	
6 7	Very Good	1.50	93.32%	54.67%	71.10%	85.55%	2.2500	1.1250	
7 8	Strong	1.75	85.99%	61.84%	78.40%	89.20%	3.0625	1.5313	
8 9	Very Strong	2.00	97.72%	68.27%	84.30%	92.15%	4.0000	2.0000	
9 10	Excellent	2.25	98.78%	73.84%	90.98%	95.04%	5.0625	2.5313	
10 11	Excellent	2.50	99.38%	78.87%	94.20%	97.10%	6.2500	3.1250	
11 12	Excellent	2.75	99.70%	83.09%	97.14%	98.57%	7.5625	3.7813	
12 13	Superior	3.00	99.57%	86.64%	98.91%	99.48%	9.0000	4.5000	
Validation Scores:		5.30	5.42	5.84	5.82	5.74	5.75		
Average Validation Score: 5.66 or Good									

"A PD Validation Framework for Basel II Internal Ratings-Based Systems," Maurice P. Joseph, Commonwealth Bank of Australia, 2005

42

Keep it simple!

Backtesting PD at Level 2

- Is there a sufficient number of rating grades?
 - Relation to masterscale
 - Impact on regulatory capital
- Are credit characteristics of borrowers in the same grade sufficiently homogeneous?
- Enough grades to allow for accurate and consistent estimation of loss characteristics per grade?
- Ratings provide **ordinal ranking** of risk
 - Check whether default rates are properly ranked through the grades
 - $DR(A) < DR(B) < DR(C)$
- Ratings provide cardinal measures of risk
 - For example, “BB” credits have an average PD of about 1.5%; “B” credits have average PD of 1.84%.

43

Backtesting at Level 2

- Compare estimated PDs versus realized DRs
- Test statistics:
 - Binomial test
 - Hosmer-Lemeshow test
 - Vasicek one-factor model
 - Normal test
- Complications:
 - Not enough defaults
 - Correlation between defaults
 - Decide on significance level
- Use as early warning indicators
- Impact of Risk Rating Philosophy
 - TTC versus PIT PDs

44

Brier Score

- The Brier score is defined as

$$\frac{1}{n} \sum_{i=1}^n (\hat{PD}_i - \theta_i)^2$$

whereby n is the number of obligors, \hat{PD}_i the forecast PD, and θ_i is 1 if obligor i defaults and 0 otherwise

- The Brier score is always bounded between 0 and 1 and lower values indicate better discrimination ability.

45

The Binomial Test

- Null hypothesis H_0 : the PD of a rating category is correct
- Alternative hypothesis H_A : the PD of a rating category is underestimated
- Assumption: default events per rating category are independent!
- Given a confidence level, α (for example, 99%), H_0 is rejected if the number of defaulters k in the rating category is greater than or equal to k^* which is obtained as follows:

$$k^* = \min \{k \mid \sum_{i=k}^n \binom{n}{i} \hat{PD}^i (1-\hat{PD})^{n-i} \leq 1-\alpha\}$$

- Use Normal Approximation (CLT): for large n, $n\hat{PD} > 5$, and $n(1-\hat{PD}) > 5$, binomial distribution can be approximated as $N(n\hat{PD}, n\hat{PD}(1-\hat{PD}))$

46

The Binomial Test

- Hence, we have:

$$P(z \leq \frac{k^* - n \hat{PD}}{\sqrt{n \hat{PD}(1 - \hat{PD})}}) = \alpha, \text{ with } z \text{ following a standard normal distribution}$$

- The critical value can then be obtained as follows:

$$k^* = N^{-1}(\alpha) \sqrt{n \hat{PD}(1 - \hat{PD})} + n \hat{PD}$$

with $N^{-1}(\alpha)$ the inverse standard normal distribution.

- In terms of a maximum observed default rate p^* , we have

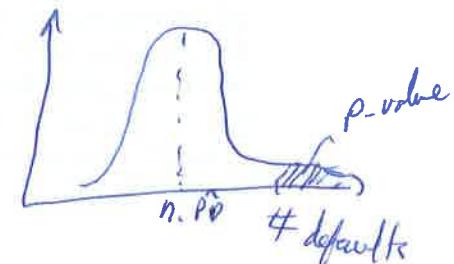
$$p^* = N^{-1}(\alpha) \sqrt{\frac{\hat{PD}(1 - \hat{PD})}{n}} + \hat{PD}$$

- Summarizing: reject H_0 at significance level α , if the observed DR is higher than p^*
- Binomial test assumes defaults are uncorrelated!
- If correlation present, higher probability to erroneously reject H_0 (type I error); use as early-warning system!

47

$$H_0: \hat{PD} = PD$$

$$H_A: \hat{PD} < PD$$

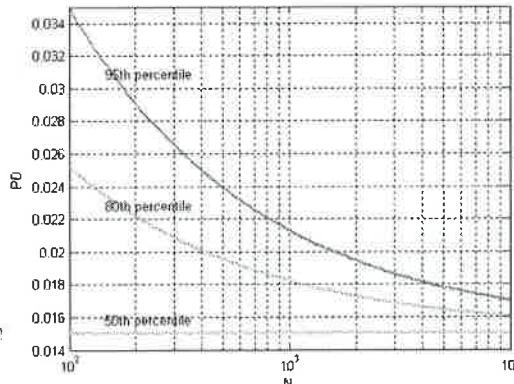


The Binomial Test

Values of the confidence interval as a function of the number of observations N (for a reference PD of 1.5%).

The width of the confidence interval decreases with growing number of observations, which makes it interesting to perform backtests on an aggregated level.

48



(Van Gestel 2005)

Binomial Test: Significance Levels

"For example, if a Binomial test is used, AIs can set tolerance limits at confidence levels of 95% and 99.9%. Deviations of the forecast PD from the realized default rates below a confidence level of 95% should not be regarded as significant and remedial actions may not be needed. Deviations at a confidence level higher than 99.9% should be regarded as significant and the PD must be revised upward immediately. Deviations which are significant at confidence levels between 95% and 99.9% should be put on a watch list, and upward revisions to the PD should be made if the deviations persist."

(Hong Kong Monetary Authority, February 2006)

49

Extensions for the Binomial Test

Take into account default correlation as follows:

$$z = \frac{DR - \hat{PD}}{\sqrt{\frac{\hat{PD}(1-\hat{PD})}{n(1-\rho^2)}}} \sim N(0, 1)$$

z-statistic becomes smaller and less conservative compared to no correlation. Hence, ignoring correlations gives more conservative tests!

50

Example Traffic Light Implementation (Van Gestel 2005)

PD backtest based on binomial test; DR from Moody's statistics used as example. The PD is estimated as the average DR.

2000

30%

PD	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av
	0.26%	0.17%	0.42%	0.53%	0.54%	1.36%	2.46%	5.76%	8.76%	20.89%	3.05%
DR	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av
1993	0.00%	0.00%	0.00%	0.83%	0.00%	0.78%	3.24%	6.04%	11.29%	28.57%	3.24%
1994	0.00%	0.00%	0.00%	0.00%	0.00%	0.59%	1.89%	3.75%	7.95%	5.13%	1.88%
1995	0.00%	0.00%	0.00%	0.00%	0.00%	1.78%	4.35%	6.42%	4.06%	11.57%	2.51%
1996	0.00%	0.00%	0.00%	0.00%	0.00%	0.00%	1.17%	0.00%	3.28%	13.99%	0.78%
1997	0.00%	0.00%	0.00%	0.00%	0.00%	0.47%	0.00%	1.54%	7.22%	14.87%	1.41%
1998	0.00%	0.31%	0.00%	0.00%	0.62%	1.12%	2.11%	7.55%	5.62%	15.06%	2.83%
1999	0.00%	0.00%	0.34%	0.47%	0.00%	2.00%	3.28%	6.21%	9.63%	20.44%	3.35%
2000	0.28%	0.00%	0.97%	0.94%	0.63%	1.04%	3.26%	4.10%	10.88%	19.65%	3.01%
2001	0.27%	0.27%	0.00%	0.51%	1.33%	1.33%	3.19%	11.07%	11.36%	21.46%	3.48%
2002	1.25%	0.12%	1.75%	1.53%	1.11%	1.58%	2.00%	6.81%	6.80%	20.11%	3.79%
Av	0.26%	0.17%	0.42%	0.53%	0.54%	1.38%	2.46%	5.76%	8.76%	20.89%	3.05%

61

The Hosmer-Lemeshow Test

- The Hosmer-Lemeshow test can be used to test several rating categories simultaneously.
- It also assumes independence of defaults.
- The test statistic is defined as follows:

$$T = \sum_{i=1}^k \frac{(n_i \hat{PD}_i - \theta_i)^2}{n_i \hat{PD}_i (1 - \hat{PD}_i)}$$

n_i =number of obligors with rating i, \hat{PD}_i is the estimated PD of rating i, θ_i is the number of observed defaulters with rating i, k is the number of ratings

- Background: binomial distribution per rating, approximate by standard normal, square to get $\chi^2(1)$, then sum across all the ratings to get a $\chi^2(k)$ for k ratings
- T converges towards a χ^2 distribution with k degrees of freedom.
- Compute p-value and decide on significance.
- Type I error underestimated, when correlation is present!

62

The Normal Test

- The normal test is a multi-period test of correctness of a default probability forecast for a single rating category.
- H_0 : none of the true probabilities of default in the years $t=1, \dots, T$, is greater than its corresponding forecast \hat{PD}_t
 H_A : not H_0
- Reject H_0 at confidence level α if:

$$\frac{\sum_{t=1}^T (DR_t - \hat{PD}_t)}{\sqrt{T\tau}} > z_\alpha, \quad \tau^2 = \frac{1}{T-1} \left(\sum_{t=1}^T (DR_t - \hat{PD}_t)^2 - \frac{1}{T} \left(\sum_{t=1}^T (DR_t - \hat{PD}_t) \right)^2 \right)$$

z_α is the standard normal α -quantile, (for example,
 $z_{0.99} \approx 2.33$)

53

Vasicek One-Factor Model

- Backtest at portfolio level using Vasicek one-factor model
- Define α^* to be the percentage of defaults that will not be exceeded at the 99.9% confidence level,
 $P(X \leq \alpha^*) = 0.999$

$$\alpha^* = N \left[\frac{N^{-1}(\hat{PD}) + \sqrt{\rho} N^{-1}(0.999)}{\sqrt{1-\rho}} \right]$$

- 99.9% confidence interval thus becomes $[0 ; \alpha^*]$
- Allows to take into account correlated defaults (via asset correlation)
- Use Basel II correlations (or half to be more conservative!)
- Assumes infinitely granular portfolio
 - Use Monte Carlo simulation for finite (small) samples

54

Vasicek Test for Small Samples

- (1) Generate a random variable $\eta \sim N(0, 1)$ representing a factor common to all asset returns (e.g. overall state of economy)
- (2) Generate a vector of n random variables $\epsilon_i \sim N(0, 1)$

$$\begin{bmatrix} A_1 \\ \vdots \\ A_n \end{bmatrix} = \sqrt{\rho}\eta + \sqrt{1-\rho} \begin{bmatrix} \epsilon_1 \\ \vdots \\ \epsilon_n \end{bmatrix} \quad (17)$$

- (3) Define the return thresholds that lead to default as $T = \Phi^{-1}(PD)$.
- (4) Calculate the average default rate (DR) in the simulated sample as:

$$DR = \frac{\sum_{i=1}^n I(A_i \leq T)}{n}, \quad (18)$$

where the indicator function I is equal to 1 if $A_i \leq T$ and zero otherwise.

- Repeat simulation multiple times to get PD distribution and confidence interval
- Use Basel II correlations (or half to be more conservative)

65

Data Aggregation

- Assume portfolio with N obligors and n risk classes
- Approximately N/n observations per risk class
- More risk classes makes backtesting more difficult
- Aggregate data to improve significance of backtesting
 - Merge risk classes with low number of observations, for example, AA+ & AA & AA- into one overall rating class AA
 - Full portfolio or important segments

66

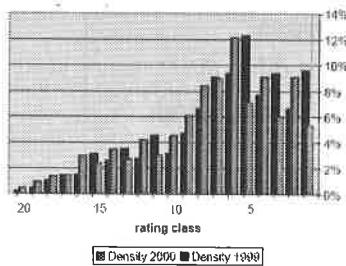
Implications of Risk Rating Philosophy on Backtesting

- Point in Time (PIT) ratings
 - Backtesting should find that realized default rates are close to forecast PD
 - PIT PDs validated against 12-month default rates
- Through the Cycle (TTC) ratings
 - Backtesting should find that realized default rates vary around forecast PD (rising in downturns and falling in upturns)
 - TTC PDs validated against cycle average default rates

57

Backtesting PD at Level 2

- Change in portfolio distribution of default rates



- Is the change due to:
 - Cyclical effect (PIT rating)
 - Systematic changes in population (input variables)

58

Validating Inputs and Outputs of the Regulatory and Economic Capital Allocation Process, Fritz, Deutsche Bank, 2003

Example Traffic Light Indicator Dashboard

Level 2: Calibration	Quantitative			
		Binomial	Not significant at 95% level	Significant at 95% but not at 99% level
		Hosmer-Lemeshow	Not significant at 95% level	Significant at 95% but not at 99% level
		Vasicek	Not significant at 95% level	Significant at 95% but not at 99% level
	Qualitative	Normal	Not significant at 95% level	Significant at 95% but not at 99% level
		Portfolio distribution	Minor shift	Moderate shift
		Difference	Correct	Overestimation
		Portfolio stability	Minor migrations	Moderate migrations
				Major migrations

59

Example Traffic Light Indicator Dashboard

Level 1: Discrimination	Quantitative			
		AR difference with reference model	< 5%	Between 5% and 10%
		AUC difference with reference model	< 2,5%	Between 2,5% and 5%
		Model significance	p-value < 0.01	p-value between 0.01 and 0.10
	Qualitative	Preprocessing (missing values, outliers)	Considered	Partially considered
		Coefficient signs	All as expected	Minor exceptions
		Number of overrides	Minor	Moderate
		Documentation	Sufficient	Minor issues
				Major issues

60

Backtesting LGD and EAD Calibration

	Rating 1	Rating 2	...	Rating n	Non-rated	Average
Estimated LGD						
Actual LGD year t						
Actual LGD year t-1						
Actual LGD year t-2						
...						
Average LGD period 1						
Average LGD period 2						

Use a *t*-test as follows:

$$t = \frac{\frac{1}{n} \sum_{i=1}^n LGD_i - LGD^*}{\frac{s}{\sqrt{n}}}$$

whereby LGD^* represents the estimated LGD and *t* follows a Student's *t*-distribution with $n-1$ degrees of freedom.

65

Benchmarking

- Comparison of internal risk estimates with external estimates and/or models
- Example benchmarking quantities:
 - credit scores, ratings, calibrated risk measurements (PD/LGD/EAD), migration matrices,
 - ...
- Example benchmarking partners:
 - credit bureaus, rating agencies, data poolers, internal experts, ...

66

Benchmarking Problems

- Unknown quality of external ratings
- Different methodologies/processes/portfolio compositions
- Take into account rating philosophy (PIT versus TTC)
- Different default/loss definition
- Different LGD weighting scheme, discount factor, collection policy, ...
- Legal constraints (for example, banking secrecy)
- Endogeneity of credit risk (?)
 - Dependent on credit culture/credit process
- Cherrypicking

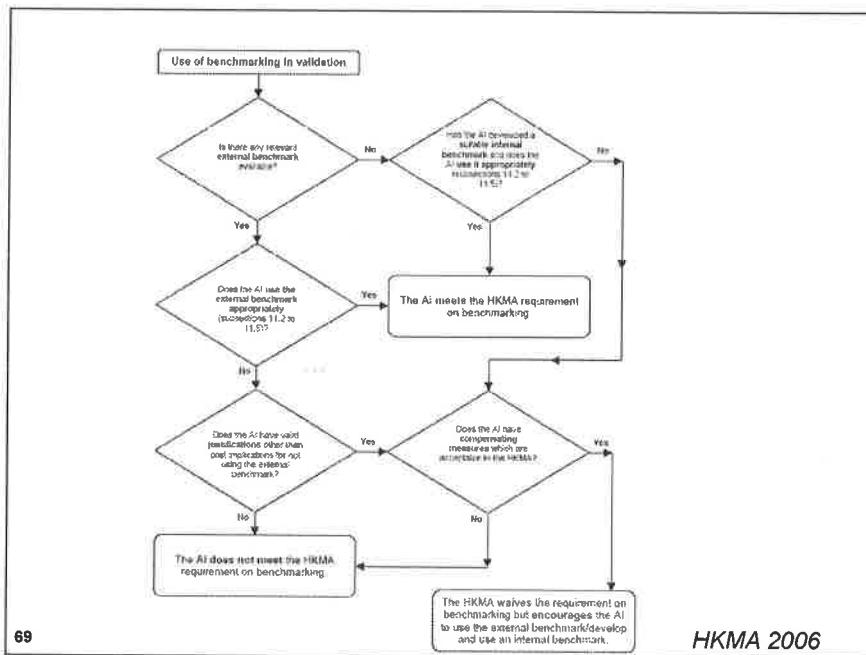
67

Benchmarking

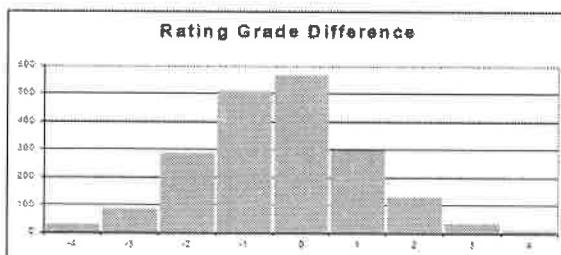
“Where a relevant external benchmark is not available (e.g., PD for SME and retail exposures, LGD, and EAD), an AI should develop an internal benchmark. For example, to benchmark against a model-based rating system, an AI might employ internal rating reviewers to re-rate a sample of credit on an expert-judgment basis.”
 (HKMA 2006)

- Develop second internal model
 - either expert-based or statistical
- Adopt champion-challenger approach
- Idea of internal benchmarking also discussed in Federal Register

68



Benchmarking: Use of Histograms



Beaumont
2005

Van Gestel 2005

Rating comparison	<-2 notches	-2 notches	-1 notch	0 notch	+1 notch	+2 notches	>+2 notches	Nobs
Performance year t								
Performance year t-1								
Performance year t-2								
Performance year t-3								
...								
Av performance period 1								
Av performance period 2								
...								

70

Spearman's Rank-Order Correlation

- Measures the degree to which a monotonic relationship exists between the scores or ratings provided by an internal rating system and those from a benchmark
- Compute numeric rankings by assigning 1 to lowest score or ranking, 2 to second lowest score, ...
- Take the average in case of tied scores or ratings
- Spearman's ρ_S can then be computed as:

$$\rho_S = 1 - \frac{6 \sum_{i=1}^n d_i^2}{n(n^2 - 1)},$$

whereby n is the number of obligors, and d_i the difference between the rankings

- Always ranges between -1 (perfect disagreement) and +1 (perfect agreement)

71

score	rating 1	rating 2	rank	rank	d	ρ_S
110	2.5	7.0	2.5	3.0	-0.5	-0.5
100	1	7.0	1.5	7.0	+1	+1
110	2.5	7.0	1.5	7.0	+1	+1
120	4	8.0	3	8.0	0	0
130	5	9.5	5	9.5	0	0

Kendall's Tau

- Assume we want to compare the estimates of institution X to those provided by benchmark Y for n debtors.
- Two debtors are said to be concordant if the debtor who is higher rated by X is also higher rated (scored) by Y, and are discordant if the debtor higher rated (scored) by X is lower rated by Y. The two cases are neither concordant nor discordant if they are tied on X or Y or both.
- Let A represent the number of concordant pairs and B represent the number of discordant pairs:

$$\text{Kendall's tau} = \frac{A - B}{\frac{1}{2} n(n - 1)}$$

The denominator represents the number of possible pairs for n debtors.

- Kendall's tau is 1 for perfect agreement, -1 for perfect disagreement
- Increasing values imply increasing agreement between the ratings or scores.

72

Goodman-Kruskal Gamma

- Note that Kendall's tau statistic assumes no ties.
- Goodman-Kruskal Gamma is defined as follows:

$$\text{Gamma} = \frac{A - B}{A + B}$$

where A represents the number of concordant pairs, and B the number of discordant pairs.

- Goodman-Kruskal Gamma ignores all tied pairs, tied on either X or Y.
- Goodman-Kruskal Gamma is +1 if there are no discordant pairs, -1 if there are no concordant pairs, and 0 if there are equal numbers of concordant and discordant pairs.

73

Example Benchmarking Exercise

By comparing the rating criteria of its internal rating system with those of Moody's, an institution concludes that 50% of the borrowers assigned to its rating grade B would have Moody's ratings "Baa1", 25% "A3", and 25% "Ba1". In the past five years, average annual default rates of these Moody's ratings were 3%, 2%, and 4% respectively. The benchmark PD of rating grade B can then be estimated as follows:

$$50\% \times 3\% + 25\% \times 2\% + 25\% \times 4\% = 3\%$$

Compare with own PD and decide!

74

Qualitative Validation

- Use testing
- Data quality
- Model design
- Documentation
- Corporate governance and management oversight

75

Use Test

- “Internal ratings and default and loss estimates must play an essential role in the credit approval, risk management, internal capital allocations, and corporate governance functions of banks using the IRB approach.” (par. 444, Basel II Accord)
- “The systems and processes used by a bank for risk-based capital purposes must be consistent with the bank’s internal risk management processes and management information reporting systems.” (Federal Register)
- Credit pricing, credit approval, economic capital calculation, ...

76

continued...

Use Test

- Essential does not necessarily mean exclusive or primary.
- Three conditions to meet use test requirement (FSA)
 - **Consistency:** information IRB estimates (PD/LGD/EAD) are based on is consistent with internal lending standards and policies
 - **Use of all relevant information:** any relevant information used in internal lending standards and policies is also used in calculating IRB estimates
 - **Disclosure:** if differences exist between calculation of IRB estimates and internal purposes, it must be documented and reasonableness demonstrated

77

Use Test

- Application scoring PD versus Basel PD
 - Time window:
 - One year for Basel II versus 18 months for application scoring
 - Default definition
 - 90 days for Basel II
 - defaulter versus bad payer
- Downturn LGD
 - “Firms can use different LGDs for business purposes to those used for regulation and not fail the use test, provided that the rationale for their use and differences/transformation to capital numbers is understood.” (FSA, 2005)
 - For example, for economic capital calculation, IFRS provisions (different discount rates), IAS 39 (no indirect costs), ...
- Document and demonstrate reasonableness to supervisor!

78

Data Quality

- “Data input into the IRB systems is accurate, complete, and appropriate.” (FSA CP 05/03, par. 7.16)
- Accuracy
 - Do the inputs measure what they are supposed to measure (for example, data accuracy scorecard in FSA CP189)
 - Data entry errors, measurement errors, and outliers are all signs of bad data accuracy
- Completeness
 - Observations with missing values can only be removed if sound justifications can be given
 - “While missing data for some fields or records may be inevitable, institutions should attempt to minimize their occurrence and aim to reduce them over time.” (CEBS, CP10, par. 297, 2005)

79

Data Quality

- Timeliness
 - Use recent data
 - Data should be updated at least annually
 - Higher updating frequencies for riskier borrowers
- Appropriateness
 - No biases or unjustified data truncation
- Data definition
 - Define data in appropriate way (for example, ratios)
 - Value “0” for zero or missing value (not both!)
- For internal, external, and expert data!
- Develop a data quality program

80

Survey: Data Quality for Credit Risk Analytics

50+ banks participating world-wide

Focus on credit risk analytics

Key findings:

- Most banks indicated that between **10–20 percent** of their data suffer from data quality problems.
- **Manual data entry** is one of the key problems.
- Diversity of data sources and **consistent** corporate wide data representation the main challenges for data quality.
- **Regulatory compliance** is the key motive to improve data quality.

Moges, Lemahieu, Baesens, 2011

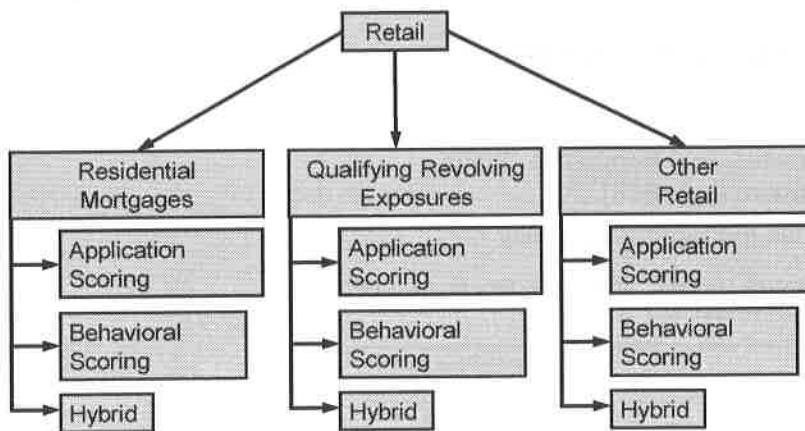
81

Data Quality Criteria (Moges, Lemahieu, Baesens, 2011)

Cat.	DQ dimensions	Definitions
Intrinsic	Accuracy (AC)	The extent to which data are certified, error-free, correct, flawless and reliable
	Objectivity (OBI)	The extent to which data are unbiased, unprejudiced, based on facts and impartial
	Reputation (REP)	The extent to which data are highly regarded in terms of its sources or content
	Completeness (COM)	The extent to which data are not missing and covers the needs of the tasks and is of sufficient breadth and depth of the task at hand
Contextual	Appropriateness (APM)	The extent to which the volume of information is appropriate for the task at hand
	Value-added (VAD)	The extent to which data are beneficial and provides advantages from its use
	Relevance (REL)	The extent to which data are applicable and helpful for the task at hand
	Timeliness (TIM)	The extent to which data are sufficiently up-to-date for the task at hand
	Actionable (ACT)	The extent to which data is ready for use
Representation	Interpretable (INT)	The extent to which data are in appropriate languages, symbols, and the definitions are clear
	Easily-understandable (EUS)	The extent to which data are easily comprehended
	Representational-consistent (RC)	The extent to which data are continuously presented in some format
	Concise-represented (CR)	The extent to which data is compactly represented, well-presented, well-organized, and well-formatted
Access	Alignment (AL)	The extent to which data is reasonable
	Accessibility (ACC)	The extent to which data is available, or easily and swiftly retrievable
	Security (SEC)	The extent to which access to data is restricted appropriately to maintain its security
	Traceability (TRA)	The extent to which data is traceable to the source

82

Impact of Segmentation on Validation



87

Impact of Segmentation on Validation

- Segmentation usually done for
 - Statistical reasons (variable interactions)
 - Operational (application versus behavioral scoring)
 - Strategic
- Statistical segmentation versus expert Segmentation
- Already enforced by Basel II
- Beware not to over-segment!
- Effects
 - Lower number of defaults per segment
 - More efforts needed for validation/backtesting/benchmarking
- Compare benefit of segmentation with cost!

88

Pillar 3 Reporting

- Encourage market discipline by developing a set of disclosure requirements that will allow market participants to assess key pieces of information on the scope, application, capital, risk exposures, risk assessment processes, and hence the capital adequacy of the institution. (par. 808)
- External reporting of risk management quality

Template 3.III.5 : Commercial and industrial portfolio: number of defaults for all PD grades at the time of default in the foundation approach for period t"

	Performing grades PD					Non-performing grades PD	
	grade 1	grade 2	grade 3	grade n		grade x	etc
Estimated PD							
Actual PD							

89

Van Gestel 2005

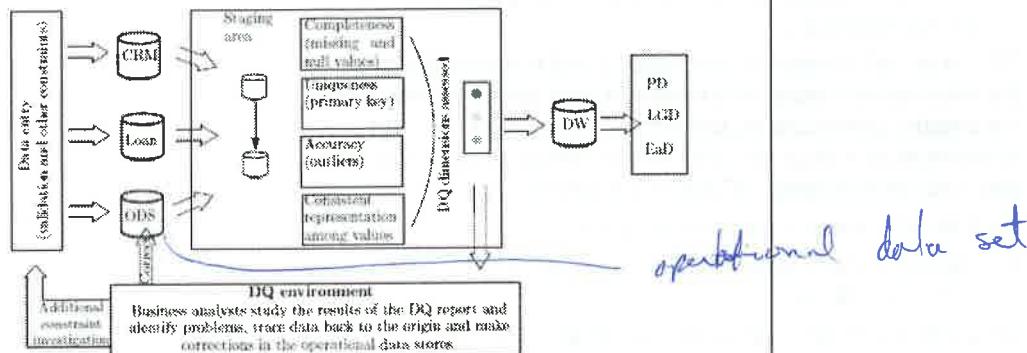
Pillar 3 Reporting

(I) PD Grade (%)	(II) Exposure	(III) Weighted Average % LGD(for advanced banks only)	(IV) Weighted Average Maturity (for advanced banks only)	(V) Value of exposures defaulting in the last year	(VI) % Default Rate
PD1					
PD2					
PD3					
PD4					
PD5					
PD6					
Default1					
Default2					

Van Gestel 2005

90

Data Quality ICT Architecture (Moges, Lemahieu, Baesens, 2011)



83

Model Design

- When was the model designed and by whom?
- What is the perimeter of the model?
 - Counterparty types, geographical region, industry sectors, ...
- What are the strengths and weaknesses of the model?
- What data was used to build the model? How was the sample constructed? What is the time horizon of the sample? Which default definition was adopted?
- How were the ratings defined?
- Is human judgment used and how?
- ...

84

Documentation

- See paragraphs 418–421 of the Basel II Accord.
- Documentation should be transparent and comprehensive
 - Federal Register: A bank's advanced systems should be transparent.
- "Documentation should encompass, but is not limited to, the internal risk rating and segmentation systems, risk parameter quantification processes, data collection and maintenance processes, and model design, assumptions, and validation results." (Federal Register)
- Both for internal and external models.
- Use document management systems with appropriate versioning facilities!
- Documentation test: can a new team use existing documentation to continue development or production of an IRB system?

85

Corporate Governance and Management Oversight

- Involvement of the board of directors and senior management in the implementation and validation process
- Senior management is responsible for sound governance of the IRB framework
- Board and senior management should have a general understanding of the IRB systems
- Should demonstrate active involvement on an on-going basis, assign clear responsibilities, and put into place organizational procedures and policies that will allow the proper and sound implementation and validation of the IRB systems
- Outcome of the validation exercise must be communicated to senior management and, if needed, accompanied by appropriate response

86

Chapter 15 Low Default Portfolios

15.1 Low Default Portfolios..... 15-3

Example: Oversampling

Original data			Oversampled data		
ID	Variables	Class	ID	Variables	Class
1		Defaulter	1		Defaulter
2		Non-Defaulter	1		Defaulter
3		Non-Defaulter	2		Non-Defaulter
4		Defaulter	3		Non-Defaulter
5		Non-Defaulter	4		Defaulter
6		Non-Defaulter	4		Defaulter
7		Non-Defaulter	5		Non-Defaulter
8		Non-Defaulter	6		Non-Defaulter
9		Defaulter	7		Non-Defaulter
10		Non-Defaulter	8		Non-Defaulter
Train			9		Defaulter
			10		Non-Defaulter
Test					

5

Low Default Portfolios (LDP)

- No formal definition in accord of LDP
 - 20 defaults? (Benjamin et al., FSA, 2006)
- Low default versus small in size (low data portfolio)
- For example, exposures to sovereigns, banks, project finance, large corporations, specialized lending, new products, ...
- Lack of sufficient statistical data and resulting difficulty in backtesting might exclude LDPs from IRB treatment (for example, paragraph 449 and 501)
- Historical averages are inappropriate
- Credit risk might be underestimated because of data scarcity
- Substantial portion of bank's assets might consist of LDPs!

6

Low Default Portfolios

- Views of the Basel Committee Accord Implementation Group's Validation Subgroup (AIGV)
 - "... LDPs should not, by their very nature, automatically be excluded from IRB treatment"
 - "... an additional set of rules or principles specifically applying to LDPs is **neither necessary nor desirable**"
 - "... relatively sparse data might require increased reliance on alternative data sources and data-enhancing tools for quantification and alternative techniques for validation"
 - "... LDPs should not be considered or treated as conceptually different from other portfolios"
- "We believe that it should be possible to include firm's LDPs in the IRB approach." (FSA, CP 05/03, par. 1.35)

7

Low Default Portfolios: Modeling Approaches

- Pooling of data with other banks or market participants; use of external data
- Aggregate subportfolios with similar risk characteristics to increase default history
- Combine rating categories and analyze PDs for the combined category in a manner consistent with paragraphs 404-405 of the Basel Accord (for corporations, sovereigns and banks)
- Use upper bound on PD estimate as input to capital requirement formulas
- Infer PD estimates with a horizon of more than one year and then annualize the resulting figure
- Use lowest non-default rating as proxy for default (still need to do calibration of ratings to PD consistent with Basel II definition)

8

continued...

17.1 Neural Networks

Overview

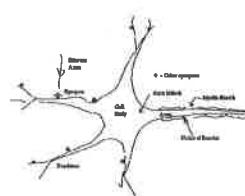
- Background
- The Multilayer Perceptron (MLP)
- Transfer functions
- Data preprocessing
- Weight learning
- Overfitting
- Architecture selection
- Input selection
- Two-stage modeling

3

Background: Two Views

Biological view

- Neural networks are mathematical representations inspired by the functioning of the human brain
- Neurons: Brain cells
- Nucleus (at the center)
- Dendrites provide inputs
- Axons send outputs



Statistical view

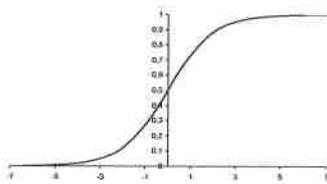
- Neural networks are generalizations of existing statistical models
- For example, linear and logistic regression

4

Linear Regression for Classification

- $y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$
- Use OLS (Ordinary Least Squares regression) to estimate β_i
- Data set: $y=1$ if class 1; $y=0$ if class 0
- Statistical problems
 - Residuals are not normally distributed (binomial)
 - Residuals have unequal variances
- Prediction can be > 1 and < 0 !
- Regression discriminant analysis (related to Fisher discriminant)
- Using a bounding function to limit the model outcome between 0 and 1, e.g., the logistic transform:

$$f(z) = \frac{1}{1 + e^{-z}}$$



6

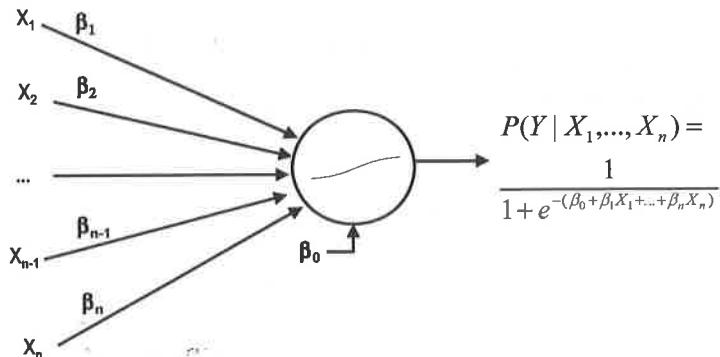
A Brief Review of Logistic Regression

$$P(Y=1 | X_1, \dots, X_n) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n)}} = \frac{e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}{1 + e^{\beta_0 + \beta_1 X_1 + \dots + \beta_n X_n}}$$

- Logistic transformation ensures that $0 \leq P(Y=1 | X_1, \dots, X_n) \leq 1$
- No distributional assumptions on the independent variables!
- Maximum likelihood estimation with Newton-Raphson optimization
- Linear in the log odds, hence the logistic regression classifier assumes a linear decision boundary
- Generalized Additive Model (GAM) because of transforming the independent variables
(For example, categorization, weights of evidence, ...)

6

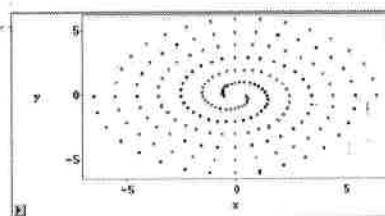
Neural Network Representation of a Logistic Regression Model



7

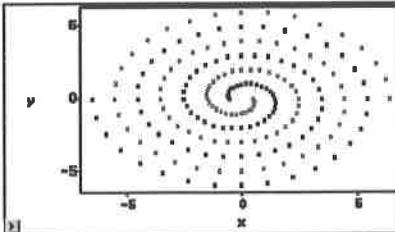
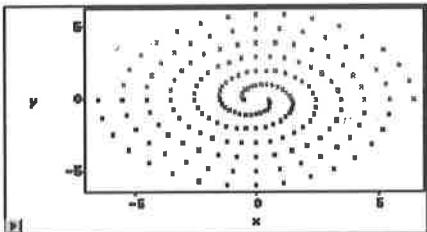
Linear versus Nonlinear Decision Boundary

The SAS System



Logistic Regression

Neural Network (40 hidden neurons)



8

A Review of Linear Regression

Linear model

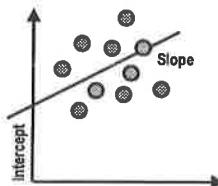
$$Y = \beta_0 + \beta_1 x_1 + \dots + \beta_n x_n = X\beta$$

Ordinary least squares (OLS) regression yields the following closed form analytical formula:

$$\beta = (X^T X)^{-1} X^T Y$$

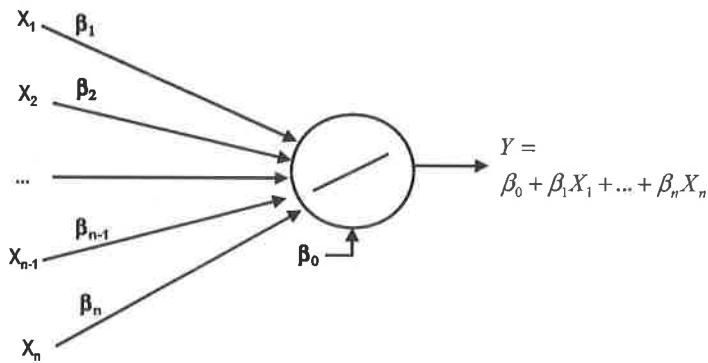
Statistical tests to decide on relevance of variables

Confidence intervals

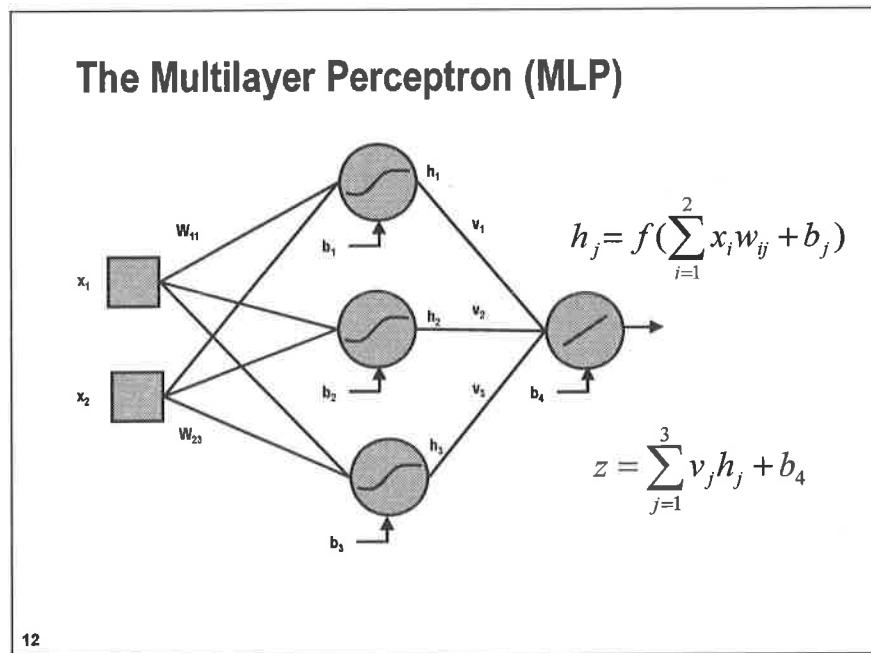
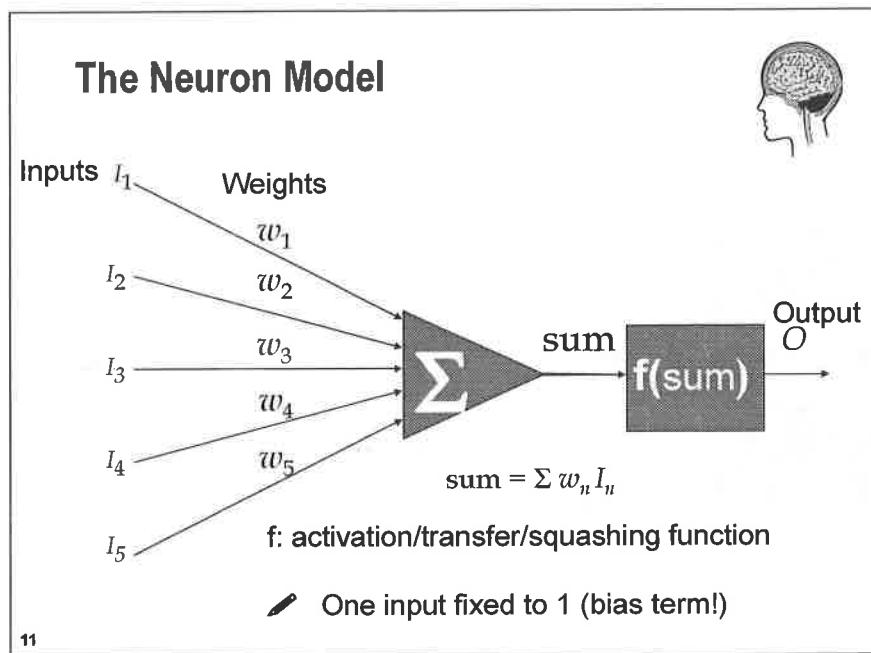


9

Neural Network Representation of Linear Regression

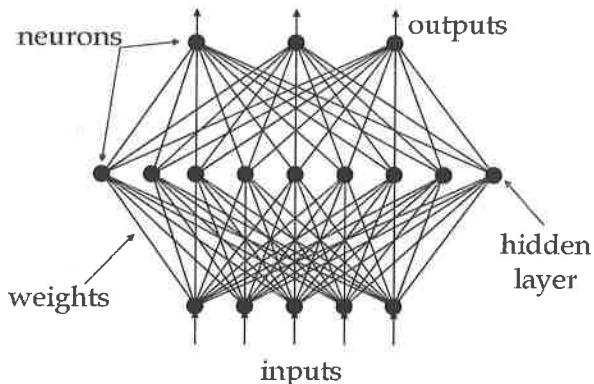


10



The Multilayer Perceptron

Organize neurons into layers!



13

The Multilayer Perceptron (MLP)

Every neuron has a bias input (intercept in linear regression)

Combination functions

- Linear (default in SAS): linear combination of incoming values and weights
- Additive: adds all incoming values without using weights or biases
- Equal slopes: same weights for each neuron per layer with different biases

Multiple hidden layers possible

- Universal approximation property with one hidden layer
- No feedback connections (recurrent networks)!

Skip layer neural network: direct connections between inputs and output

MLPs are the most common type of neural network for supervised prediction (for example, credit scoring, churn prediction, response modeling, ...)

If k inputs, h hidden neurons (one hidden layer), one output, then $(k+1)h+h+1$ weight parameters for fully connected MLP

14

Terminology: Neural Networks versus Statistics

Neural Networks	Statistics
Learning/Training	Optimization, Estimation
Bias	Intercept
Weight	Parameter estimate
Epoch	Iteration Step
Pruning	Model reduction, input selection
Architecture	Model

15

Activation Function f

logistic (sigmoid)

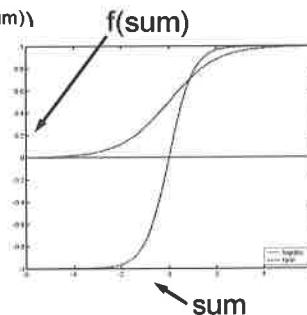
- $f(\text{sum}) = 1/(1+e^{(-\text{sum})})$
- between 0 and 1

hyperbolic tangent (tanh)

- $f(\text{sum}) = (e^{(\text{sum})} - e^{(-\text{sum})}) / (e^{(\text{sum})} + e^{(-\text{sum})})$
- between -1 and 1
- same as $2 \cdot \text{logistic}(2 \cdot \text{sum}) - 1$
- default in SAS for hidden units

linear

- $f(\text{sum}) = \text{sum}$
- between $-\infty$ and $+\infty$
- no transformation



continued...

16

Activation Function f

Exponential

- $f(\text{sum}) = e^{\text{sum}}$
- between 0 and $+\infty$ (positive targets)

Radial Basis Function (RBF Networks)

- Gaussian activation functions
- Lots of theory but not many practical data mining applications

Other activation functions in SAS: arctan and Elliot
(less frequently used, but available in SAS)

Practical advice:

- **For classification:** use tanh in hidden and logistic in output layer
- **For regression:** use tanh in hidden and linear/tanh in output layer

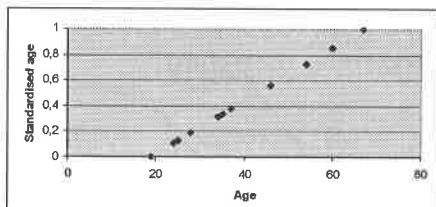
17

Data Preprocessing

Continuous variables

- Standardize to the range [0; 1], [-1; 1] using a linear transformation
- Use the z-scores

$$z_i = \frac{x_i - \bar{X}}{S_X}$$



Categorical variables

- Use 0/1 dummy coding
- Use weights of evidence (WOE) coding
- If too many, do categorization/coarse classification (using, for example, chi-squared or decision tree analysis)

18

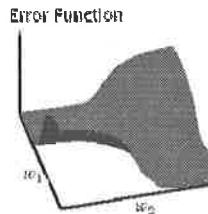
Training Neural Networks

For simple statistical models (for example, linear regression), there exists closed-form analytical formulas for the optimum parameter estimates (cf. supra).

For nonlinear models like neural networks, the parameter estimates need to be determined numerically, using an iterative algorithm.

The error function defines a surface in weight space.

Error function measures the difference between the target and predicted values.



19

Error Functions

For regression

- Mean squared error (MSE)

$$E = \frac{1}{2} \sum_{i=1}^N (y(\mathbf{x}_i; \mathbf{w}) - t_i)^2$$

with $y(\mathbf{x}_i; \mathbf{w})$ the output of the network for observation i and t_i the true output value (target) of observation i

For classification

- Cross entropy:

$$E = - \sum_{i=1}^N \sum_{k=1}^c (t_{ik} \ln(y_{ik}(\mathbf{x}_i; \mathbf{w})))$$

with t_{ik} a 0/1 variable indicating class membership and y_{ik} the output of the network for observation i and output class k (1-of- c coding)

- Cross entropy for binary classification:

$$E = - \sum_{i=1}^N (t_i \ln(y(\mathbf{x}_i; \mathbf{w})) + (1-t_i) \ln(1-y(\mathbf{x}_i; \mathbf{w})))$$

20

Error Functions in SAS Enterprise Miner

- For interval targets, the default is least squares minimization.
- For binary classification, the cross-entropy error is used (same as Bernouilli deviance).
- For multiclass classification, the multiple Bernouilli deviance is used.

Practical advice:

- Default works fine!
- Even using mean squared error (MSE) for classification can be OK!

21

Optimizing the Weights

Error function defines a parameter surface

Need to find minimum of error function using characteristics of the surface

Algorithm overview

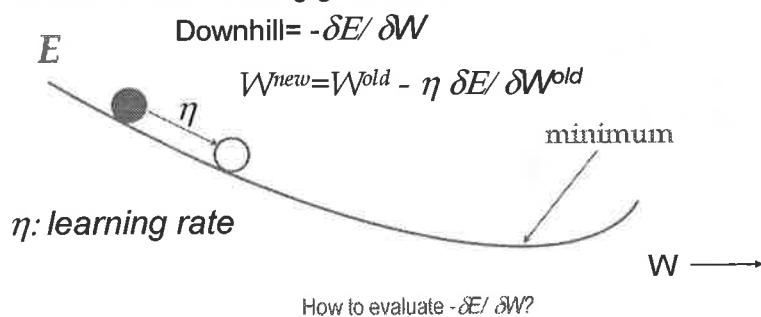
- Start with weights w^0
- Recursively update w until convergence

22

Gradient Descent Learning

The error E is a function of the weights and the training data.

Minimize error E using gradient descent.

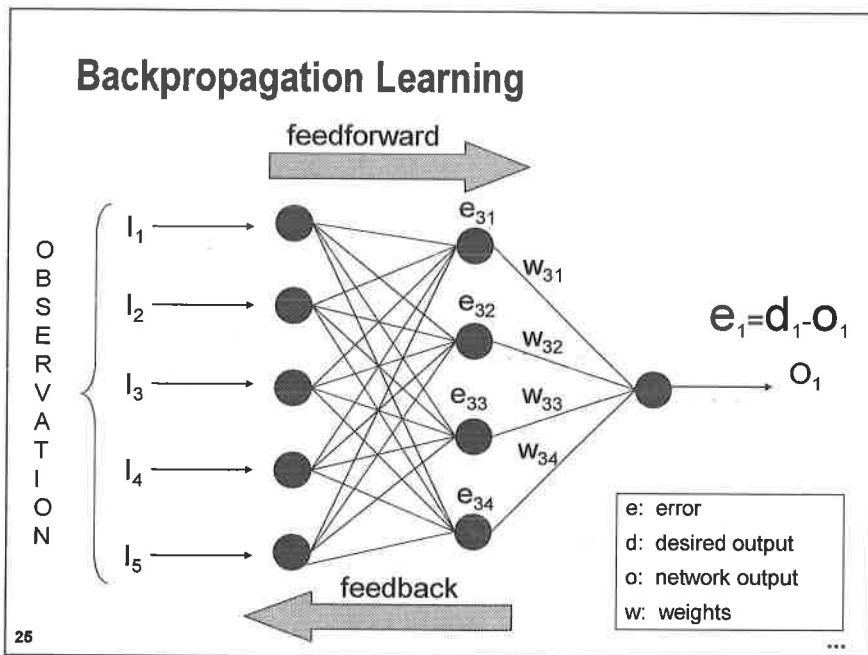


23

Backpropagation Learning

- Step 1: Initialize the weights to small values
- Step 2: Propagate an input pattern through the network
- Step 3: Compute the errors at the output layer
- Step 4: Propagate the errors backward to the preceding layers
- Step 5: Update all weights
- Step 6: Go back to step 2, until convergence is achieved

24



Backpropagation: Characteristics

Online (incremental) learning

- Update the weights after reading each observation

Batch learning

- Update the weights after reading the entire data set
- ✍ 1 epoch=1 run of all observations through the network

Learning rate η

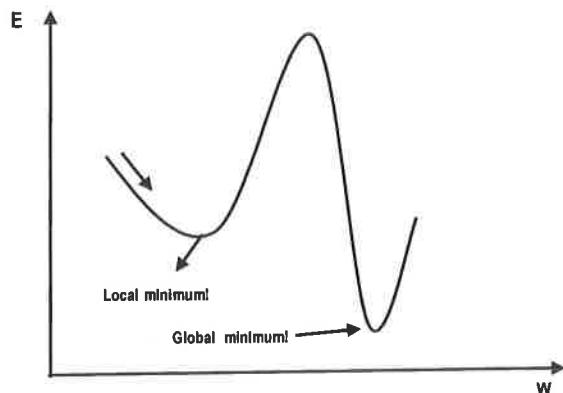
- Too high: oscillations, algorithm will diverge
- Too small: slow progress
- Adaptive learning rate!

Momentum parameter α

- $0 \leq \alpha \leq 1$
- Current step direction is smoothed with the previous step

Local minima

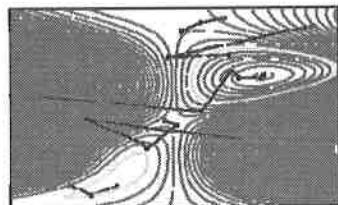
Local Minima



27

Preliminary Training

- Good starting values for the weights are essential for getting good solutions.
- Preliminary training uses a small number of random starting weights and takes a few iterations (20 by default) from each.
- Use the best of the final values as the new starting value.



28

Neural Network Modelling, William Potts, SAS, 2004

Advanced Training Algorithms

- Advanced nonlinear optimization algorithms
- Take into account Hessian
 - Matrix of second-order partial derivatives of the error function with respect to the weights
 - Curvature of the error surface
- Newton based methods
- Conjugate gradient
- Levenberg-Marquardt
- Again, the default settings in SAS generally work well across a wide range of nonlinear modeling problems!

29

Convergence Criteria

- Objective function shows no progress
- The weight parameter estimates stop changing substantially
- The gradient is close to zero
- Use a validation set (see later)

30

Learning versus Overfitting

Successful Learning:

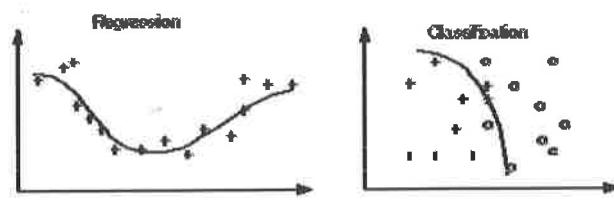
- Recognize data outside the training set, that is, data in an independent test set.

Overfitting ('Memorization')

- Each data set is characterized by noise (idiosyncrasies) due to incorrect entries, human errors, irrationalities, noisy sensors,
- A model that is too complex (for example, decision tree with too many nodes, neural network with too many hidden neurons) might fit the noise, not just the signal, leading to overfitting.

31

Learning versus Overfitting



overfitting:

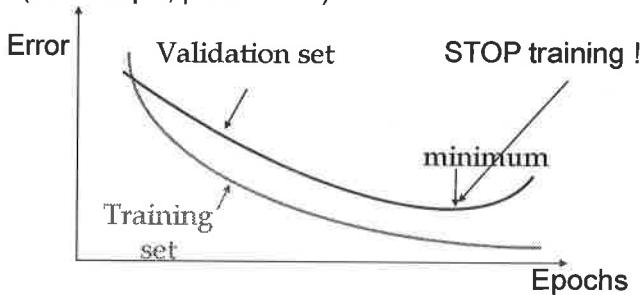
true model



32

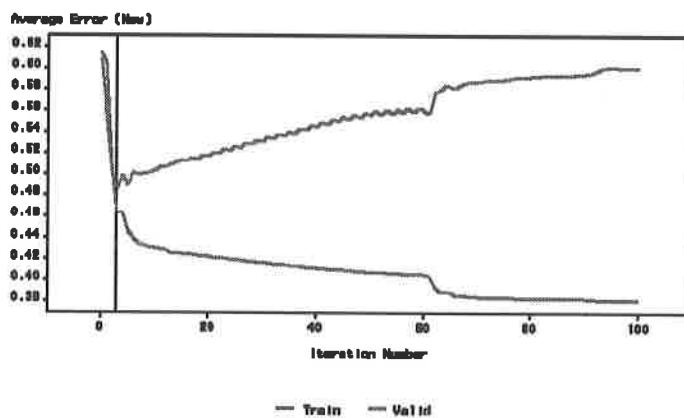
Avoiding Overfitting Using Early Stopping

- Set aside validation set (for example, 30% of observations)
- Start from oversized network (enough hidden neurons)
- Use training set to estimate weights and validation set to decide when to stop adjusting the weights when overfitting occurs ('early stopping')
- Evaluation metric can be different than objective function (for example, profit based)



33

Avoiding Overfitting Using Early Stopping



34

Avoiding Overfitting Using Regularization

- Bartlett (1997) demonstrated that generalization depends more on the size of the weights than the number of weights. A large network with small weights acts like a smaller, less complex network.
- Large weights cause the sigmoids to saturate.
- Thus, restraining the weights should prevent a bumpy overfitted surface.
- However, it might also prevent the model from adapting to true features of the data.
- Penalize large weights in the objective function!

35

Avoiding Overfitting Using Regularization

Objective function = Error function + $\lambda ||w||^2$

The decay (shrinkage, smoothing) parameter λ controls the severity of the penalty

Trade-off

- Setting λ too low might cause overfitting
- Setting λ too high might cause underfitting

Tune λ using a separate validation set

Ridge/Lasso regression in statistics

Compared to early stopping

- No need for validation data set, hence no loss of data

Separate weight regularization terms for different weight groups

- Automatic Relevance Determination (ARD)
- See, for example, Baesens et al., 2002.

36

Designing the Neural Network Architecture

Number of hidden layers

- One, because of universal approximation property (continuous function)
- Two hidden layers for discontinuous functions (very rare)

Transfer functions

- Cf. infra

Number of hidden neurons

- Grid Search
- Sequential Network Construction (SNC)
- Cascade Correlation
- Bayesian Methods
- Genetic Algorithms
- **Practical advice:** use grid search idea

37

Grid Search

1. Split data into a training set, validation set, and test set.
2. Vary the number of hidden neurons from one to 10 using a predefined step size (for example, 1).
3. Train the neural networks on the training set and measure their performance on the validation set.
 - Try out a number of neural networks each time to avoid bad local minima.
4. Choose the number of hidden neurons with optimal (average) validation set performance.
5. Measure the final performance on the independent test set.

38

Input Selection

In statistical methods: use (standardized) magnitude of the coefficients of the inputs to evaluate their importance.

More complicated in neural networks because each input is associated with h input-to-hidden layer weights and h hidden-to-output weights.

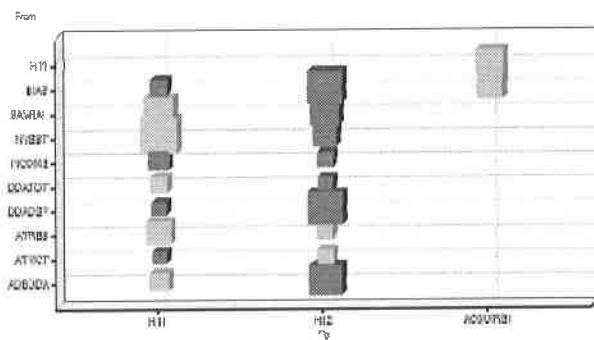
However, if all input-to-hidden weights for an input are zero (or close), the input is clearly not important.

Pruning algorithm

- Train the neural network.
- Prune the input where the input-to-hidden weights are closest to zero.
- Retrain the network (might be starting from previous weight set to speed up convergence).
- If the predictive power increases (or stays the same) then repeat; if not, reconnect the input and stop.

39

Hinton Diagrams



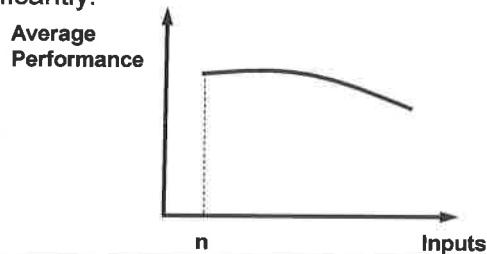
**Size of square indicates size of weight
Color of square indicates sign of weight**

Neural Network Modelling, William Potts, SAS, 2004

40

Brute Force Input Selection

- Start from training k networks with all n inputs (for example, k=5).
- Remove each input in turn and train k networks.
- Remove the input giving the best average performance (for example, maximal Area under ROC, minimal Mean Squared Error, ...).
- Repeat this procedure until performance decreases significantly.



41

Two-Stage Model

- Estimate an easy to understand model first (for example, linear regression, logistic regression).
- Use a neural network to predict the errors made by the simple model using the same set of predictors.
- Combine both models in an additive way!
 - Target=linreg(X_1, X_2, \dots, X_N)+NN(X_1, X_2, \dots, X_N)
 - Score=logreg(X_1, X_2, \dots, X_N)+NN(X_1, X_2, \dots, X_N)
- Ideal balance between model interpretability and model performance.
- Do not estimate in 1 multivariate setup, otherwise too much weight can be put on nonlinear part!

42

More Information on Neural Networks

- *Neural Networks for Pattern Recognition*,
Chris Bishop, Oxford University Press, 1999
- *Pattern Recognition and Neural Networks*,
Brian D. Ripley, Cambridge University Press, 1996
- *Introduction to Artificial Neural Networks*,
Jacek M. Zurada, PWS Publishing Company, 1992
- Journals
 - IEEE Transactions on Neural Networks
 - Neural Computation
 - Neural Networks
 - Neural Processing Letters



Chapter 18 New Techniques for PD/LGD Modeling: Support Vector Machines

18.1 Support Vector Machines 18-3

18.1 Support Vector Machines

Problems with Neural Networks

- Multimodal objective function
 - Multiple local minima
- Highly parameterized
 - How to choose hidden layers?
 - How to choose hidden neurons?
 - How to set regularization parameter?

3

Linear Programming

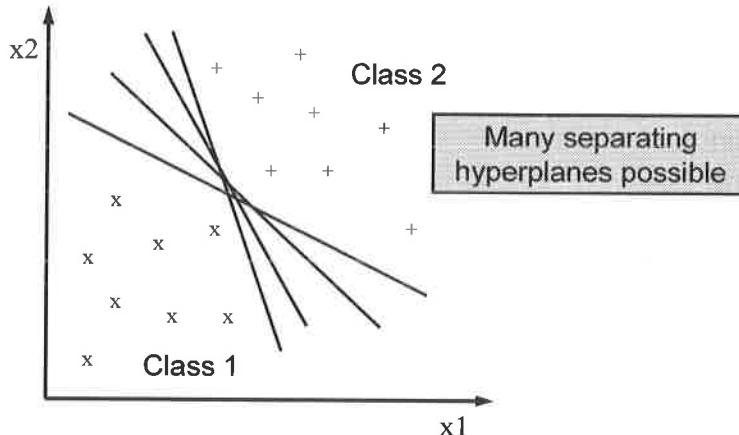
- Mangasarian (1965)
- Minimize the sum of the absolute values of the deviations (MSD)

$$\begin{aligned} & \min e_1 + e_2 + \dots + e_{n_g} + \dots + e_{n_b} \\ & \text{subject to} \\ & w_1 x_{i1} + w_2 x_{i2} + \dots + w_n x_{in} \geq c - e_i, \quad 1 \leq i \leq n_g, \\ & w_1 x_{i1} + w_2 x_{i2} + \dots + w_n x_{in} \leq c + e_i, \quad n_g + 1 \leq i \leq n_g + n_b, \\ & e_i \geq 0. \end{aligned}$$

- Use fixed cutoff c , but experiment both with negative and positive c (Freed and Glover 1986)

4

The Linear Separable Case

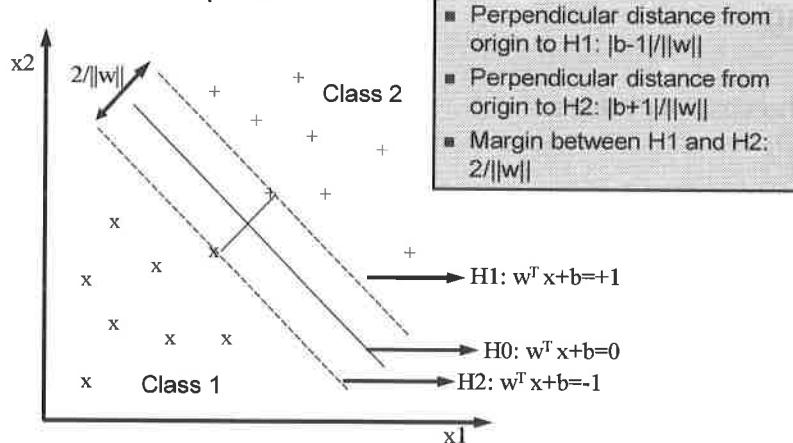


5

continued...

The Linear Separable Case

Consider the hyperplane, which maximizes the distance to the nearest points.



6

The Linear Separable Case

- Large margin separating hyperplane (Vapnik 1964)
- Given a training set $\{x_k, y_k\}_{k=1}^N, x_k \in \Re^n, y_k \in \Re$, where $y_k \in \{-1,+1\}$
- Assume

$$\begin{cases} w^T x_k + b \geq +1, & \text{if } y_k = +1 \\ w^T x_k + b \leq -1, & \text{if } y_k = -1 \end{cases}$$

- Or,
- $y_k [w^T x_k + b] \geq 1, \quad k = 1, \dots, N$
- Maximize the margin, or minimize $\frac{1}{2} \|w\|^2 = \frac{1}{2} \sum_{i=1}^n w_i^2$
- Optimization problem

$$\min \frac{1}{2} \sum_{i=1}^n w_i^2$$

$$y_k [w^T x_k + b] \geq 1, \quad k = 1, \dots, N$$

7

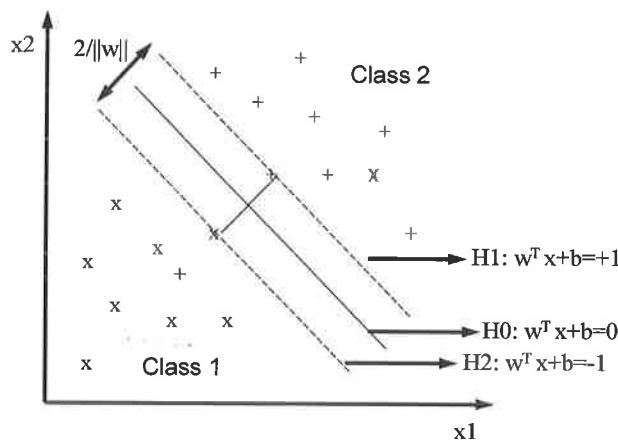
continued...

The Linear Separable Case

- The classifier then becomes: $y(x) = \text{sign}(w^T x + b)$
 - Note $\text{sign}(x) = +1$ if $x \geq 0$; -1 otherwise
- Using Lagrangian optimization, a quadratic programming (QP) problem is obtained.
- The solution of the QP problem is global.
 - convex optimization
- Training points that lie on one of the hyperplanes H_1 or H_2 are called *support vectors*.

8

The Nonseparable Case



9

The Nonseparable Case

- Allow for errors by introducing slack variables in the inequalities.

$$\begin{aligned} y_k [w^T x_k + b] &\geq 1 - \varepsilon_k, \quad k = 1, \dots, N \\ \varepsilon_k &\geq 0 \end{aligned}$$

- The optimization problem then becomes

$$\min_{w, \varepsilon} I(w, \varepsilon) = \frac{1}{2} \sum_{i=1}^n w_i^2 + C \sum_{k=1}^N \varepsilon_k$$

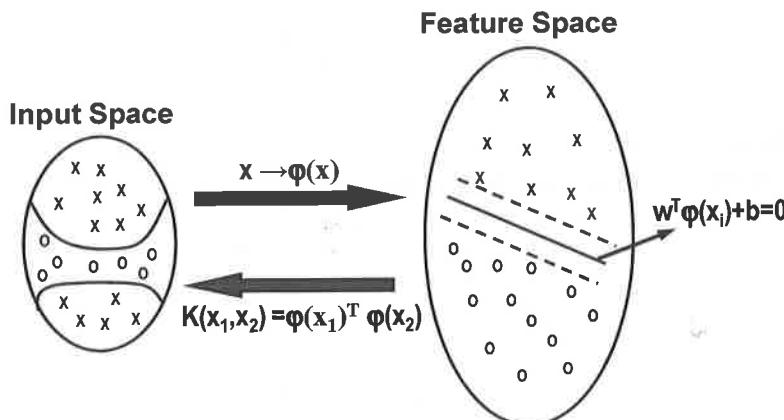
subject to

$$\begin{aligned} y_k [w^T x_k + b] &\geq 1 - \varepsilon_k, \quad k = 1, \dots, N \\ \varepsilon_k &\geq 0 \end{aligned}$$

- C is a user-defined parameter, higher C corresponding to higher penalty for errors.
- Again, a QP problem is obtained after Lagrangian optimization.

10

The Nonlinear SVM Classifier



11

The Nonlinear SVM Classifier

- Map data from input space to high-dimensional feature space and construct linear hyperplanes in feature space:

$$\begin{cases} w^T \phi(x_k) + b \geq +1, & \text{if } y_k = +1 \\ w^T \phi(x_k) + b \leq -1, & \text{if } y_k = -1 \end{cases}$$

- The optimization problem then becomes

$$\begin{aligned} \min_{w, \varepsilon} I(w, \varepsilon) &= \frac{1}{2} \sum_{i=1}^n w_i^2 + C \sum_{k=1}^N \varepsilon_k \\ y_k [w^T \phi(x_k) + b] &\geq 1 - \varepsilon_k, \quad k = 1, \dots, N \\ \varepsilon_k &\geq 0 \end{aligned}$$

12

continued...

The Nonlinear SVM Classifier

- Construct the Lagrangian:

$$L(w, b, \varepsilon; \alpha, v) = I(w, \varepsilon_k) - \sum_{k=1}^N \alpha_k \{y_k [w^T \varphi(x_k) + b] - 1 + \varepsilon_k\} - \sum_{k=1}^N v_k \varepsilon_k$$

with Lagrange multipliers $\alpha_k \geq 0, v_k \geq 0$

- Solution given by the saddle point of the Lagrangian:

$$\max_{\alpha} \min_{w, b, \varepsilon} L(w, b, \varepsilon; \alpha, v)$$

- This yields:

$$\frac{\partial L}{\partial w} = 0 \rightarrow w = \sum_{k=1}^N \alpha_k y_k \varphi(x_k)$$

$$\frac{\partial L}{\partial b} = 0 \rightarrow \sum_{k=1}^N \alpha_k y_k = 0$$

$$\frac{\partial L}{\partial \varepsilon_k} = 0 \rightarrow 0 \leq \alpha_k \leq C, \quad k = 1, \dots, N$$

13

continued...

The Nonlinear SVM Classifier

- The dual QP problem becomes:

$$\max_{\alpha} Q(\alpha) = -\frac{1}{2} \sum_{k, l=1}^N y_k y_l K(x_k^T x_l) \alpha_k \alpha_l + \sum_{k=1}^N \alpha_k$$

$$\sum_{k=1}^N \alpha_k y_k = 0$$

$$0 \leq \alpha_k \leq C, \quad k = 1, \dots, N$$

- The kernel function is defined through the Mercer theorem: $K(x_k, x_l) = \varphi(x_k)^T \varphi(x_l)$
- w and $\varphi(x_k)$ are not calculated.

14

continued...

The Nonlinear SVM Classifier

- The nonlinear SVM classifier becomes:

$$y(x) = \text{sign} \left[\sum_{k=1}^N \alpha_k y_k K(x, x_k) + b \right]$$

with α_k positive real constants, b real constant.

- Nonzero α_k are support vectors.

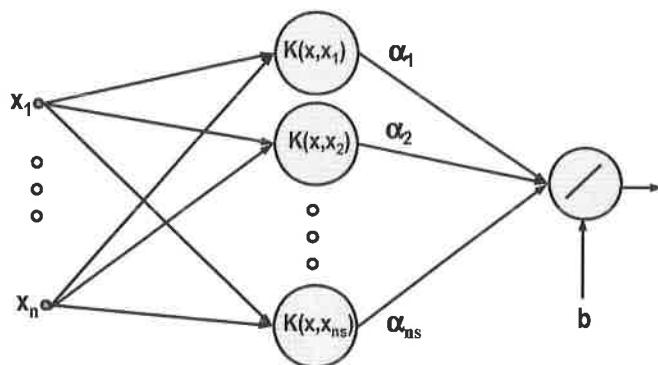
16

Kernel Functions

- $K(x, x_k) = x_k^T x$ (linear SVM)
- $K(x, x_k) = (x_k^T x + 1)^d$ (polynomial SVM of degree d)
- $K(x, x_k) = \exp\{-||x - x_k||^2/\sigma^2\}$ (RBF SVM)
- $K(x, x_k) = \tanh(\kappa x_k^T x + \theta)$ (MLP SVM)
- The Mercer condition holds for all σ values in the RBF case but not for all κ and θ values in the MLP case.
- For the RBF and MLP kernels, the number of support vectors corresponds to the number of hidden neurons.

16

Neural Network Interpretation of SVM Classifier



Number of hidden neurons determined automatically

17

The Polynomial Kernel

- Suppose $K(x,y)=(x^T.y)^2$ (x and y two-dimensional vectors).
- Take a mapping to a three-dimensional space

$$\varphi(x) = \begin{pmatrix} x_1^2 \\ \sqrt{2}x_1x_2 \\ x_2^2 \end{pmatrix} \quad \varphi(y) = \begin{pmatrix} y_1^2 \\ \sqrt{2}y_1y_2 \\ y_2^2 \end{pmatrix}$$

- Hence

$$\varphi(x)^T \cdot \varphi(y) = x_1^2 y_1^2 + 2x_1 x_2 y_1 y_2 + x_2^2 y_2^2 = \left(\begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} \right)^2 = K(x, y)$$

- Note that the mapping is not unique, for example, mapping to four-dimensional space.

$$\varphi(x) = \begin{pmatrix} x_1^2 \\ x_1 x_2 \\ x_1 x_2 \\ x_2^2 \end{pmatrix}$$

18

Tuning the Hyperparameters

1. Set aside two-thirds of the data for the training/validation set and the remaining one-third for testing.
2. Starting from $i=0$, perform ten-fold cross-validation on the training/validation set for each (σ, C) combination from the initial candidate tuning sets $\sigma_0 = \{0.5, 5, 10, 15, 25, 50, 100, 250, 500\}$ and $C_0 = \{0.01, 0.05, 0.1, 0.5, 1, 5, 10, 50, 100, 500\}$.
3. Choose optimal (σ, C) from the tuning sets σ_0 and C_0 by looking at the best cross-validation performance for each (σ, C) combination.
4. If $i=i_{\max}$, go to step 5; else $i=i+1$, construct a locally refined grid $\sigma_i \times C_i$ around the optimal hyperparameters (σ, C) and go to step 3.
5. Construct the SVM classifier using the total training/validation set for the optimal choice of the tuned hyperparameters (σ, C) .
6. Assess the test set accuracy by means of the independent test set.

19

Benchmarking Study

- Studied both SVMs and LS-SVMs
- 10 publicly available binary classification data sets and 10 publicly available multiclass classification data sets
- Various domains: medicine, physics, artificial, credit scoring, sociology, ...
- Cross-validation based grid search mechanism to tune the hyperparameters (cf. supra, $i_{\max}=3$)
- RBF kernels, Linear kernels and Polynomial kernels ($d=2, \dots, 10$)
- Minimum Output Coding and One versus One coding

20

Benchmarking Study

- Compared their performance with:
 - linear discriminant analysis, quadratic discriminant analysis, logistic regression, C4.5, oneR, 1-nearest neighbor, 10-nearest neighbor, Naive Bayes (with and without kernel approximation), majority rule
- Performance was compared using paired *t*-test
- Average rankings were computed and compared using sign test
- Four months of computer time + two additional months

21

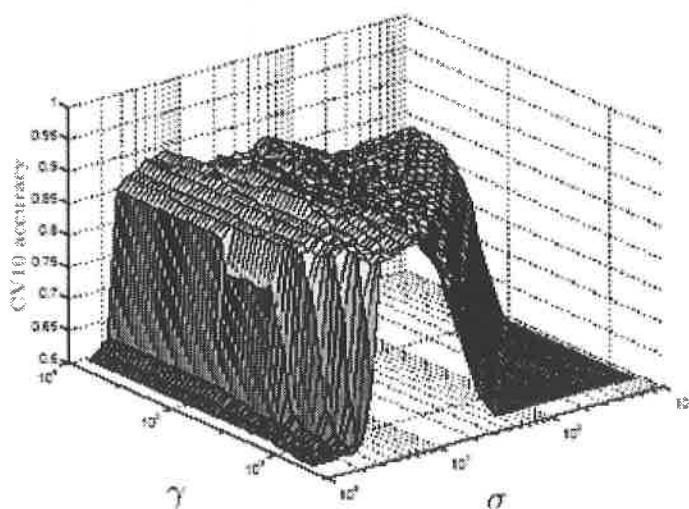
Characteristics of Data Sets

	acr	bld	gcr	hea	ion	pid	snr	ttt	vbc	adu
N_{CV}	469	230	666	180	234	312	138	638	455	33000
N_{test}	230	115	334	90	117	236	70	320	228	12222
N	699	345	1000	270	351	708	208	958	683	45222
n_{sum}	6	6	7	7	33	8	60	0	9	6
n_{rel}	8	0	13	6	0	0	0	9	0	8
n	14	6	20	13	33	8	60	9	9	14

	bal	cme	ims	iri	led	thy	usp	veh	wav	win
N_{CV}	416	982	1540	100	2000	4800	6000	504	2400	118
N_{test}	209	491	770	50	1000	2400	3298	282	1200	60
N	625	1473	2310	150	3000	7200	9298	846	3600	178
n_{sum}	4	2	18	3	0	6	256	18	19	13
n_{rel}	0	7	0	0	7	15	0	0	0	0
n	4	9	18	4	7	21	256	18	19	13
M	3	3	7	3	10	3	10	4	3	3
L_{MOG}	2	2	3	2	4	2	4	2	2	2
L_{avg}	3	3	21	3	45	3	45	6	2	3

22

Flat Maximum Effect



23

Performance for Binary Classification Data

	scr	bld	gcf	sea	ion	pid	sar	scr	scr	vbc	adu	AA	AR	Pat
Net	230	115	353	90	117	230	70	323	229	12222				
%	34	6	90	13	N/A	8	60	9	8	14				
RBF LS-SVM	87.0 (2.1)	70.2 (1.7)	76.3 (1.5)	44.7 (3.8)	29.0 (2.7)	76.8 (1.7)	70.3 (2.2)	99.8 (0.3)	90.2 (1.0)	96.7 (0.3)	88.4	3.2	0.727	
RBF LS-SVM _F	80.1 (2.0)	67.0 (2.0)	70.0 (1.0)	45.2 (3.0)	31.4 (2.7)	72.0 (2.0)	63.0 (2.4)	97.0 (0.5)	96.8 (0.7)	99.5 (0.1)	87.0	3.8	0.199	
LR LS-SVM	80.6 (2.2)	69.0 (2.2)	70.0 (2.2)	31.0 (2.8)	18.5	68.0 (2.0)	70.1 (1.8)	92.0 (0.9)	95.5 (0.6)	97.0 (0.3)	79.8	7.7	0.109	
LR LS-SVM _F	80.5 (2.2)	64.0 (2.0)	68.0 (2.0)	31.0 (2.5)	18.5	70.1 (1.8)	70.1 (1.4)	97.0 (0.5)	96.0 (0.6)	97.0 (0.3)	77.3	5.5	0.109	
PLS LS-SVM	80.5 (2.2)	62.0 (2.0)	70.0 (1.4)	33.5 (2.5)	17.0	70.0 (1.7)	70.0 (1.7)	99.5 (0.5)	99.0 (0.5)	99.5 (0.3)	84.2	4.1	0.727	
PLS LS-SVM _F	80.6 (2.2)	65.0 (2.0)	70.0 (2.0)	33.5 (2.5)	17.0	70.0 (1.8)	70.0 (1.8)	99.5 (0.5)	99.0 (0.5)	99.5 (0.3)	82.0	6.2	0.311	
RBF SVM	80.3 (1.9)	70.4 (2.2)	75.0 (1.4)	48.4 (2.4)	35.3 (1.7)	77.0 (2.0)	75.0 (2.1)	99.0 (0.5)	99.0 (0.5)	99.0 (0.3)	84.5	9.0	1.026	
LR SVM	80.7 (2.4)	67.0 (2.0)	75.4 (1.7)	35.0 (2.1)	27.0 (1.8)	77.0 (2.0)	74.0 (2.7)	97.0 (0.5)	99.5 (0.4)	97.0 (0.3)	79.8	7.0	0.623	
LSV	85.0 (2.2)	67.4 (2.2)	73.0 (2.2)	34.3 (2.0)	17.0	71.0 (2.0)	67.0 (2.0)	99.0 (0.0)	95.5 (1.1)	97.0 (0.9)	79.8	6.8	0.701	
Q2SA	80.1 (1.9)	65.2 (2.0)	70.5 (1.4)	37.0 (2.0)	20.0 (1.7)	73.0 (2.0)	73.0 (2.0)	97.0 (0.1)	91.0 (1.0)	92.0 (0.9)	79.8	2.8	0.024	
L-PLS	80.6 (2.1)	64.0 (2.0)	70.0 (2.1)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	5.8	0.109	
PSO-S	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF}	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0	77.0 (2.1)	67.0 (2.0)	97.0 (0.0)	90.0 (1.1)	92.0 (0.9)	79.8	2.8	0.024	
PSO-S _{TF} _F	85.7 (2.1)	65.0 (2.0)	70.0 (2.0)	33.0 (2.0)	16.0 </td									

24

Performance for Multiclass Classification Data

25

Conclusions

- RBF SVMs and RBF LS-SVMs yield very good classification performances compared to the other algorithms.
 - For the multiclass case, the One versus One coding scheme yielded better performance than the minimum output coding scheme.
 - Simple classification algorithms (for example, linear discriminant analysis and logistic regression) also yield satisfactory results.
 - Most data sets are only weakly nonlinear.
 - But: importance of marginal performance benefits (for example, credit scoring)

26

Support Vector Machines in SAS

```

proc dmdb data=neural.dmlspir batch out=dmdb dmdbcat=meta;
  var x y;
  class c(desc);
  target c;
run;

proc svm data=dmdb dmdbcat=meta nomonitor out=spiralout
  kernel=RBF K_PAR=0.5 c=100;
  var x y;
  target c;
run;

```

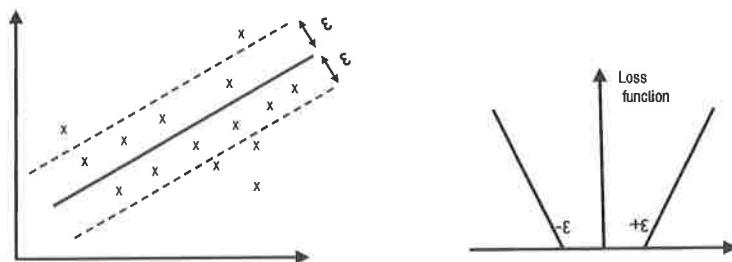
27

SVMs for Regression

Given a training set: $\{x_k, y_k\}_{k=1}^N, x_k \in \mathbb{R}^n, y_k \in \mathbb{R}$

Find a function $f(x)$ that has at most ϵ deviation from the actual targets y_i for all the training data, and is at the same time as flat as possible.

Errors lower than ϵ are tolerated (ϵ -insensitive loss regression).



28

SVMs for Regression

The SVM formulation then becomes:

$$\min \frac{1}{2} \sum_{i=1}^n w_i^2 + C \sum_{k=1}^N (\xi_k + \xi_k^*)$$

subject to

$$\begin{cases} y_k - w^T x_k - b \leq \varepsilon + \xi_k \\ w^T x_k + b - y_k \leq \varepsilon + \xi_k^* \\ \xi_k, \xi_k^* \geq 0 \end{cases}$$

The constant C determines the trade-off between the flatness of f and the amount to which deviations larger than ε are tolerated.

29

SVMs for Regression

Similar to the classification case, construct Lagrangian and solve the dual formulation which has a global minimum.

The regression function then becomes:

$$f(x) = \sum_{k=1}^N (\alpha_k - \alpha_k^*) x_k^T x + b$$

where α_i and α_i^* are the Lagrange multipliers.

Note that the SVM algorithm only depends on the dot product between the various observations, hence, you can use kernels to perform nonlinear regression using the feature space idea.

The nonlinear regression function then becomes:

$$f(x) = \sum_{k=1}^N (\alpha_k - \alpha_k^*) K(x_k, x) + b$$

Note that C and ε need to be tuned using, for example, a cross-validation procedure.

30

More Information on Support Vector Machines

- *An introduction to support vector machines*, Nello Cristianini and John Shawe-Taylor, Cambridge University Press, 2000
- *Kernel Methods for Pattern Analysis*, John Shawe-Taylor and Nello Cristianini, Cambridge University Press, 2004
- *The nature of statistical learning theory*, Vladimir Vapnik, Springer, 1995
- *Learning with Kernels*, Bernard Schölkopf and Alex Smola, MIT Press, 2002
- www.kernel-machines.org

Chapter 19 New Techniques for PD/LGD Modeling: Survival Analysis

19.1 Survival Analysis 19-3

19.1 Survival Analysis

Survival Analysis for Credit Scoring

- Traditionally, credit scoring models aim at distinguishing good customers from bad customers.
- However, the timing of the problem is also important!
- The advantages of having models that estimate when customers default are (Banasik, Crook, and Thomas 1999; Thomas, Edelman, and Crook 2002):
 - The ability to compute the profitability over a customer's lifetime and perform profit scoring.
 - These models can provide the bank with an estimate of the default levels over time, which is useful for debt provisioning
 - The estimates might help to decide upon the term of the loan
 - Changes in economic conditions can be incorporated more easily
- Can be used to predict:
 - when customers will default
 - when customers will pay back early
 - when we can expect the next cash flow in a recovery process

3

Literature Overview: General References

- *Analysis of Survival Data*, Cox and Oakes, Chapman and Hall, 1984
- *The Statistical Analysis of Failure Time Data*, Kalbfleisch and Prentice, Wiley, New York, 1980
- *Survival Analysis: A Self-Learning Text*, Kleinbaum, Springer, New York, 1996
- *SAS Survival Analysis Techniques for Medical Research*, Alan Cantor, SAS Institute Inc., 1997
- *Survival Analysis Using the SAS System (Second Edition)*, Paul D. Allison, SAS Institute Inc., 2010

4

Literature Overview: Credit Scoring References

- Narain, B., "Survival analysis and the credit granting decision," In L.C. Thomas, J.N. Crook, and D.B. Edelman, *Credit Scoring and Credit Control*, p. 109-121, Oxford University Press, 1992.
- Banasik, J., Crook, J.N., and Thomas, L.C., "Not if but when will borrowers default," *Journal of the Operational Research Society*, 50: 1185-1190, 1999.
- Stepanova, M. and Thomas, L.C., "PHAB scores: proportional hazards behavioural scores," *Journal of the Operational Research Society*, 52(9): 1007-1016, 2001.
- Stepanova, M. and Thomas, L.C., "Survival analysis methods for personal loan data," *Operations Research*, 50(2):277-289, 2002.
- Baesens, B., Van Gestel, T., Stepanova, M., and Vanthienen, J., "Neural network survival analysis for personal loan data," *Proceedings of the Eighth Conference on Credit Scoring and Credit Control (CSCCVII'2003)*, Edinburgh, Scotland, 2003.

5

Basic Concepts of Survival Analysis

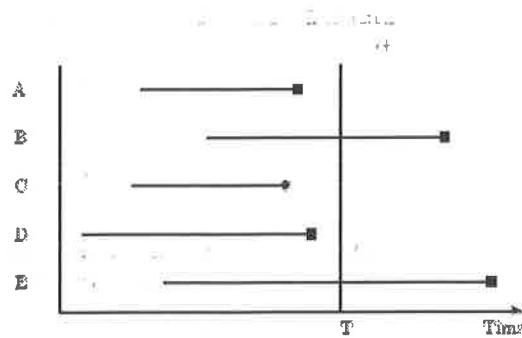
- Aim is to study occurrence and timing of events
- Event is a qualitative change that can be situated in time
 - For example, death, marriage, promotion, default, early repayment
- Originated from medicine (study of deaths)
- Two typical problems:
 - Censoring
 - Time-dependent covariates

6

Censoring

- Left-censoring versus right-censoring.
- An observation on a variable T is right-censored if all you know about T is that it is greater than some value c .
 - For example, suppose T is person's age at death, and you only know that T is > 50 ; hence, the obs is right-censored at age 50.
- An observation on a variable T is left-censored if all you know about T is that it is smaller than some value c .
 - For example, study smoking behavior and start at age 20, whereas some already begun but cannot remember
 - Less common
- Interval censoring
 - You know that $a < T < b$.

Censoring



Survival Distributions

- Event time distribution $f(t)$:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

- Cumulative distribution $F(t)$: $F(t) = P(T \leq t) = \int_0^t f(u)du$

- Survival function $S(t)$: $S(t) = 1 - F(t) = P(T > t) = \int_t^\infty f(u)du$

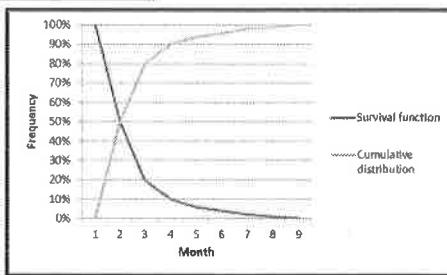
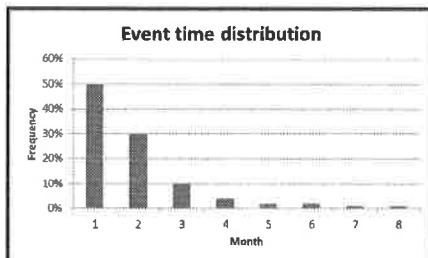
- $S(t)$ is monotone decreasing with $S(0)=1$ and $S(\infty)=0$

- The probability density function $f(t)$ is defined as

$$f(t) = \frac{dF(t)}{dt} = -\frac{dS(t)}{dt}$$

9

Survival Distributions: Examples



10

Survival Distributions

- The hazard function is defined as follows:

$$h(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t | T \geq t)}{\Delta t}$$

- The hazard function tries to quantify the instantaneous risk that an event will occur at time t , given that the individual has survived to t .
- Probability of event at time t is 0, hence look at interval between t and $t+\Delta t$.
- Only consider individuals surviving to time t , because others have already died and are no longer at risk.
- The probability is a non-decreasing function of Δt , hence, divide by Δt .
- We want the probability at exactly time t , hence, take limit for $\Delta t \rightarrow 0$.
- Hazard is a rate not a probability, hence can go from 0 (no risk) to infinity (certainty of the event).
- $h(t)$ measures the risk of the event occurring at time point t , or the expected number of events per unit of time

11

The Hazard Function

- Remember, a probability distribution can also be defined as follows:

$$f(t) = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq T < t + \Delta t)}{\Delta t}$$

- The hazard function is sometimes described as a conditional density.
- The survivor function $S(t)$, the probability density function $f(t)$, and the hazard function $h(t)$ are mathematically equivalent ways of describing a continuous probability distribution.
- The following relationships hold:

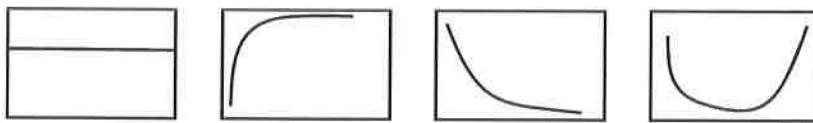
$$h(t) = \frac{f(t)}{S(t)} \quad h(t) = -\frac{d \log S(t)}{dt}$$

$$S(t) = \exp \left\{ - \int_0^t h(u) du \right\}$$

12

Example Hazard Shapes

- Constant hazard
 - Risk remains the same at all times
- Increasing
 - Natural aging
- Decreasing
 - Mortality after diagnosis of cancer decreases as patients get cured
- Convex bathtub shape
 - human mortality declines after birth and infancy, remains low for a while, and increases with elder years
 - property of some mechanical systems to either fail soon after operation, or much later, as the system ages



13

Kaplan Meier Method

- Also known as the product-limit estimator.
- Nonparametric maximum likelihood estimator for $S(t)$.
- If no censoring, the KM estimator $\hat{S}(t)$ is just the sample proportion with event times greater than t .
- If censoring, start with ordering the event times in ascending order $t_1 < t_2 < \dots < t_k$. At each time t_j , there are n_j individuals who are at risk of the event. **At risk** means that they have not undergone the event, nor have they been censored prior to t_j . Let d_j be the number of individuals who die at t_j .
- The KM estimator is then defined as follows:

$$\hat{S}(t) = \prod_{j:t_j \leq t} \left[1 - \frac{d_j}{n_j} \right] = \hat{S}(t-1) \cdot \left(1 - \frac{d_t}{n_t} \right) = \hat{S}(t-1) \cdot (1 - h(t))$$

for $t_1 \leq t \leq t_k$

- The term between brackets is the conditional probability of surviving to time t_{j+1} , given that the subject has survived to time t_j .

14

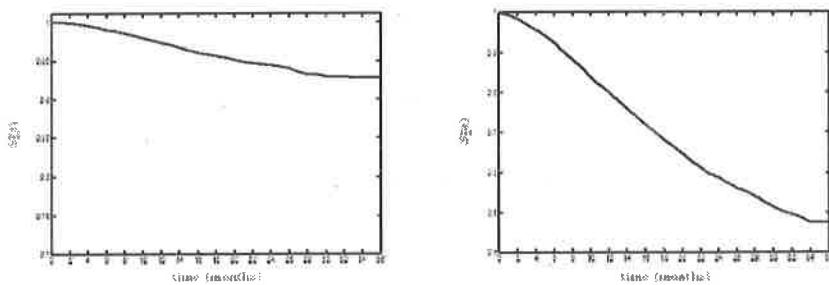
Kaplan Meier Analysis in SAS

```
proc lifetest data=credit;
  time dur*status(0);
run;
```

- PROC LIFETEST
- KM estimator is default (to be explicit: use METHOD=KM).
- The DUR variable is the time of the event, or the time of censoring.
- The status variable is the censoring indicator; value of 0 corresponds to censored observations.

15

Kaplan Meier Analysis



16

Kaplan Meier Analysis

Comparing survival curves

- H_0 : the survival curves are statistically the same
- H_1 : the survival curves are statistically different
- log-rank test (also known as the Mantel-Haenzel test),
the Wilcoxon test, and the likelihood-ratio statistic

If many unique event times, use life-table (also known as actuarial) method to group event times into intervals.

KM estimator does not account for covariates.

Test for the effect of covariates.

17

Parametric Survival Analysis Models: Exponentially Distributed Event Times

- The probability function $f(t)$ is

$$f(t) = \lambda e^{-\lambda t}$$

- The survival function then becomes

$$S(t) = e^{-\lambda t}$$

- The hazard is constant over time

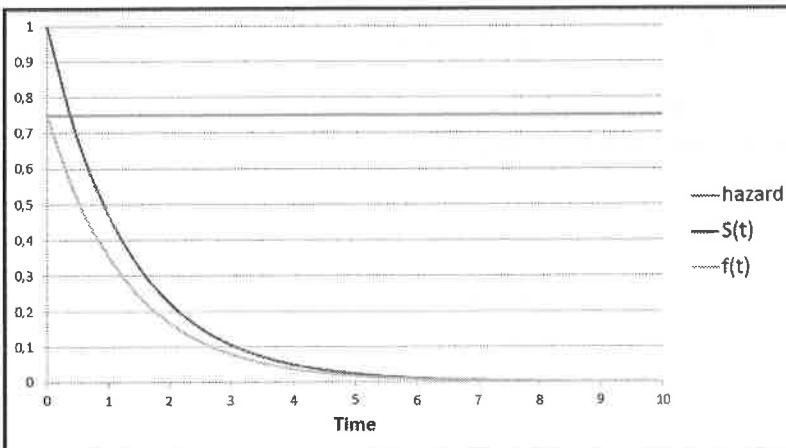
$$h(t) = \lambda$$

- When covariates are present

$$\log h(t) = \mu + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

18

Parametric Survival Analysis Models: Exponentially Distributed Event Times



19

Parametric Survival Analysis Models: Weibull Distributed Event Times

- The probability function $f(t)$ is

$$f(t) = \kappa\rho(\rho t)^{\kappa-1} \exp[-(\rho t)^\kappa]$$

- The survival function then becomes

$$S(t) = \exp[-(\rho t)^\kappa]$$

- The hazard is (not constant over time)

$$h(t) = \kappa\rho(\rho t)^{\kappa-1}$$

- When covariates are present:

$$\log h(t) = \mu + \alpha \log(t) + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n$$

20

General Parametric Survival Analysis Model

- The general model is

$$\log T_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i$$

with T_i the event time for the i^{th} individual.

- The log transformation is to ensure that predicted values of T are positive.

$$T_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i)$$

- Also known as accelerated failure time (AFT) model.
- If no censored data, estimate with ordinary least squares (OLS).
- If censored data, use maximum likelihood estimation.
- PROC LIFEREG

21

Example: Exponential Distribution

- Exponential distribution for t corresponds to constant hazard.

$$T_i = \exp(\beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \dots + \beta_n x_{in} + \varepsilon_i)$$

$$\log h(t) = \beta_0^* + \beta_1^* x_1 + \beta_2^* x_2 + \dots + \beta_n^* x_n$$

- Both models are equivalent; it can be shown mathematically that $\beta_j = -\beta_j^*$ for all j .
- If hazard is high (low), survival times are short (long).

22

Distributions for T

Distribution of T	Distribution of ϵ
Weibull	Gumbel distribution (2 par.)
Exponential	Gumbel distribution (1 par.)
Gamma	Log-gamma
Log-logistic	Logistic
Log-normal	normal

23

Maximum Likelihood Estimation

- Maximize the probability of getting the sample at hand
- Two steps
 - Construct the likelihood function for the given sample
 - Maximize the likelihood function
- Construct the likelihood function as follows:

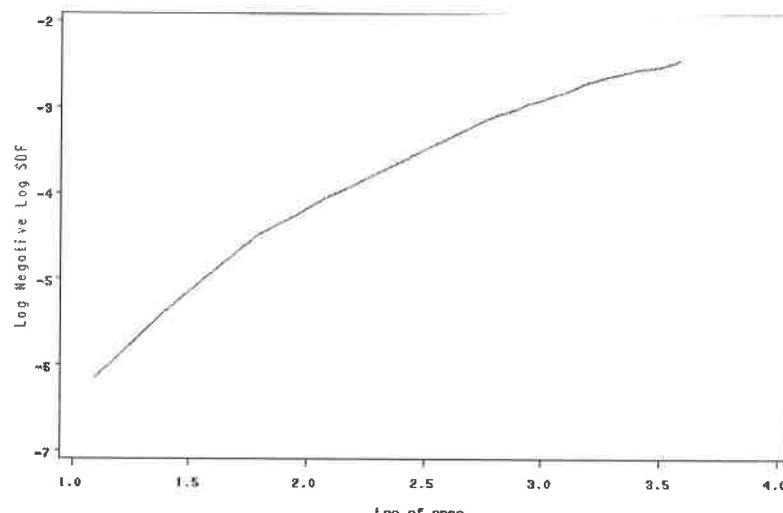
$$L = \prod_{i=1}^N f(t_i)$$

$$L = \prod_{i=1}^N f(t_i)^{\delta_i} S(t_i)^{1-\delta_i}$$

δ_i is 0 if observation is censored at t_i ; δ_i is 1 if observation dies at t_i .

24

Evaluating Model Fit Graphically



29

Evaluating Model Fit Statistically

- The likelihood ratio test statistic can be used to compare models if one model is a special case of another (nested models).
- The generalized gamma distribution is defined as follows:

$$f(t) = \frac{\beta}{\Gamma(k)\theta} \left(\frac{t}{\theta}\right)^{k\beta-1} e^{-(\frac{t}{\theta})^\beta}$$

- Suppose $\sigma = \frac{1}{\beta\sqrt{k}}$ and $\delta = \frac{1}{\sqrt{k}}$, then the Weibull, exponential, standard gamma, and log-normal models are all special versions of the generalized gamma model as follows:

$\sigma=\delta$: standard gamma

$\delta=1$: Weibull

$\sigma=1, \delta=1$: exponential

$\delta=0$: lognormal

30

continued...

Evaluating Model Fit Statistically

- Let L_{full} be the likelihood of the full model (e.g., generalized gamma) and L_{red} be the likelihood of the reduced (specialized) model (e.g., exponential).
- The chi-squared test statistic can then be computed as:

$$-2 \ln(L_{red}/L_{full})$$
- The degrees of freedom correspond to the number of reduced parameters.
 - Exponential versus Weibull: one degree of freedom
 - Exponential versus standard gamma: one degree of freedom
 - Exponential versus generalized gamma: two degrees of freedom
 - Weibull versus generalized gamma: one degree of freedom
 - Log-normal versus generalized gamma: one degree of freedom
 - Standard gamma versus generalized gamma: one degree of freedom

31

PROC LIFEREG for Parametric Survival Analysis

```
proc lifereg data=creditsurv;
  class custgend freqpaid homephon loantype
    marstat homeowns;
  model open*censor(0)=age amount curradd
    curremp custgend depchild
    freqpaid homephon insprem loantype
    marstat term homeowns
    / dist=exponential;
run;
```

- The option DIST= exponential / Weibull / lognormal / gamma / loglogistic.
- The CLASS statement automatically creates dummy variables for categorical variables (not in PROC PHREG).

32

The Proportional Hazards Model

- Basic model:

$$h(t, \mathbf{x}_i) = h_0(t) \exp\{\beta_1 x_{i1} + \dots + \beta_n x_{in}\} = h_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i\}$$
- Hazard of individual i at time t is product of:
 - a baseline hazard function $h_0(t)$ that is left unspecified (except that it can't be negative)
 - a linear function of a set of fixed covariates, which is exponentiated.
- $h_0(t)$ can be considered as hazard for individual with all covariates equal to 0.
- If x_i increases with 1, then the hazards for all t increase with $\exp(\beta_i)$, which is called the hazard ratio (HR).
- If $\beta > 0$ then $HR > 1$, $\beta < 0$ then $HR < 1$; $\beta = 0$ then $HR = 1$.
- David Cox, Regression Models and Life Tables, Journal of the Royal Statistical Society, Series B, 1972.
- Very popular.

33

continued...

The Proportional Hazards Model

- Taking logarithms yields

$$\log h(t, \mathbf{x}_i) = \alpha(t) + \beta_1 x_{i1} + \dots + \beta_n x_{in}$$
 with $\alpha(t) = \log h_0(t)$
- Hence, if $\alpha(t) = \alpha$, we get the exponential model;
 if $\alpha(t) = \alpha \log(t)$, we get the Weibull model
- However, no need to specify $\alpha(t)$ (semiparametric model)
- Taking the ratios of the hazards for individuals i and j

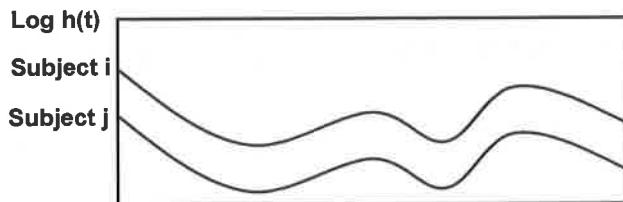
$$\frac{h_i(t)}{h_j(t)} = \exp\{\beta_1(x_{i1} - x_{j1}) + \dots + \beta_n(x_{in} - x_{jn})\} = \exp[\boldsymbol{\beta}^T (\mathbf{x}_i - \mathbf{x}_j)]$$

34

continued...

The Proportional Hazards Model

- Hazard of any individual is a fixed proportion of the hazard of any other individual (proportional hazards).
- The subjects most at risk at any one time remain the subjects most at risk at any one other time.



35

Partial Likelihood Estimation

- Suppose n individuals ($i=1,\dots,n$)
- For each individual: \mathbf{x}_i , t_i , δ_i , t_i is time of event or censoring, δ_i is 1 if uncensored, δ_i is 0 if censored observation
- Start by ranking all events of the non-censored subjects t_1, \dots, t_k
- Given the fact that one subject has event time t_i , the probability that this subject has inputs \mathbf{x}_j is then given by:

$$\frac{h(t_i, \mathbf{x}_j) \Delta t}{\sum_{l \in R(t_i)} h(t_l, \mathbf{x}_l) \Delta t} = \frac{\exp(\beta^T \mathbf{x}_j) \cdot h_0(t_i)}{\sum_{l \in R(t_i)} \exp(\beta^T \mathbf{x}_l) \cdot h_0(t_l)} = \frac{\exp(\beta^T \mathbf{x}_j)}{\sum_{l \in R(t_i)} \exp(\beta^T \mathbf{x}_l)}$$

$R(t_i)$ represents the subjects that are at risk at t_i

36

continued...

Estimating Survivor Functions

- Because

$$S(t, \mathbf{x}) = \exp\left[-\int_0^t h_0(u) \exp(\boldsymbol{\beta}^T \mathbf{x}) du\right]$$

we have,

$$S(t, \mathbf{x}) = S_0(t)^{\exp(\boldsymbol{\beta}^T \mathbf{x})}$$

with

$$S_0(t) = \exp\left(-\int_0^t h_0(u) du\right) = \exp(-\Lambda_0(t))$$

- $S_0(t)$ is the baseline survivor function (that is, survivor function for an individual whose covariates are all 0).

41

continued...

Estimating Survivor Functions

- If x_i increases with 1, the survival probabilities are raised to the power $\exp(\beta_i)$, which is the hazard ratio (HR).
 - For example, if the HR for a group with a covariate value of 1 relative to the group having the covariate value of 0 is 2.0, then for any time t , the survival probability of group 1 at t is the square of the corresponding survival probability for group 0.
- After the β parameters have been estimated, $S_0(t)$ can be estimated using a nonparametric maximum likelihood method.
- Individual survival functions can then be computed.
- Use the BASELINE statement in PROC PHREG.

42

Estimating Survivor Functions in SAS

```

data testset;
    input age amount curradd curremp;
    datalines;
29 3000 1 2
run;

proc phreg data=creditsurv;
    model open*censor(0)=age amount curradd
        curremp / ties=efron;
    baseline out=a covariates=testset
        survival=s lower=lcl upper=ucl
        / nmean;
run;

proc print data=a;
run;

```

43

Evaluating Survival Analysis Models

- Statistical significance of both the model as a whole as well as the individual covariates
- Predict time of event when $S(t) < 0,50$ and compare with real event time
- Take a snapshot of the survival probabilities at a specific time t (e.g., 12 months), compare with event time indicator and calculate $ROC(t)$
- Indicates how well the model ranks the observations for each t
- Evaluate interpretability of model by using univariate sign checks on the covariates

44

Time Dependent Covariates

- Covariates that change in value over the course of the observation (for example, behavioral scoring):

$$h(t, \mathbf{x}_i) = h_0(t) \exp\{\boldsymbol{\beta}^T \mathbf{x}_i(t)\}$$

- Note that the proportional hazards assumption no longer holds because the time-dependent inputs will change at different rates for different subjects, so the ratios of their hazards cannot remain constant.
- We can also let the β parameters vary over time:

$$h(t, \mathbf{x}_i) = h_0(t) \exp\{\boldsymbol{\beta}^T(t) \mathbf{x}_i(t)\}$$

- The partial likelihood estimation method can easily be extended to accommodate these changes in the model formulation.

45

Competing Risks

- Observation can experience any of k competing events.
- For each observation, time of the event and type of the event are observed.
- Examples
 - Customers might die because of cancer or aging
 - Early repayment versus default
- As long as a customer has not undergone any of the events, he remains at risk for any event.
- After a customer has undergone the event, he is no longer included in the population at risk for any of the other risk groups, hence he becomes censored for the other risks.
- See, for example, Crowder (2001), Kalbfleisch and Prentice (2003).

46

Drawbacks of Statistical Survival Analysis Models

- Functional relationship remains linear or some mild extension thereof
- Interaction and nonlinear terms have to be specified ad hoc
- Extreme hazards for outlying observations
- Proportional hazards assumption

47

Acknowledgements

Karen Washburn, Program Manager,
SAS Business Knowledge Series

Patsy Poole, Project Manager

Deb Bayo, Curriculum and Business Development

Dan Kelly, Statistical Training Specialist

Bob Lucas, Director, Statistical Training
and Technical Services

Susan Willard, Editor

Susan Hoggard, Production

Tonya Wells, Onsite Training Coordinator

Larry Stewart, Senior Director, Education

Lieve Goedhuys, Academic Program Manager,
SAS Belgium-Luxembourg

48

Appendix A Exercises

A.1 Exercises A-3

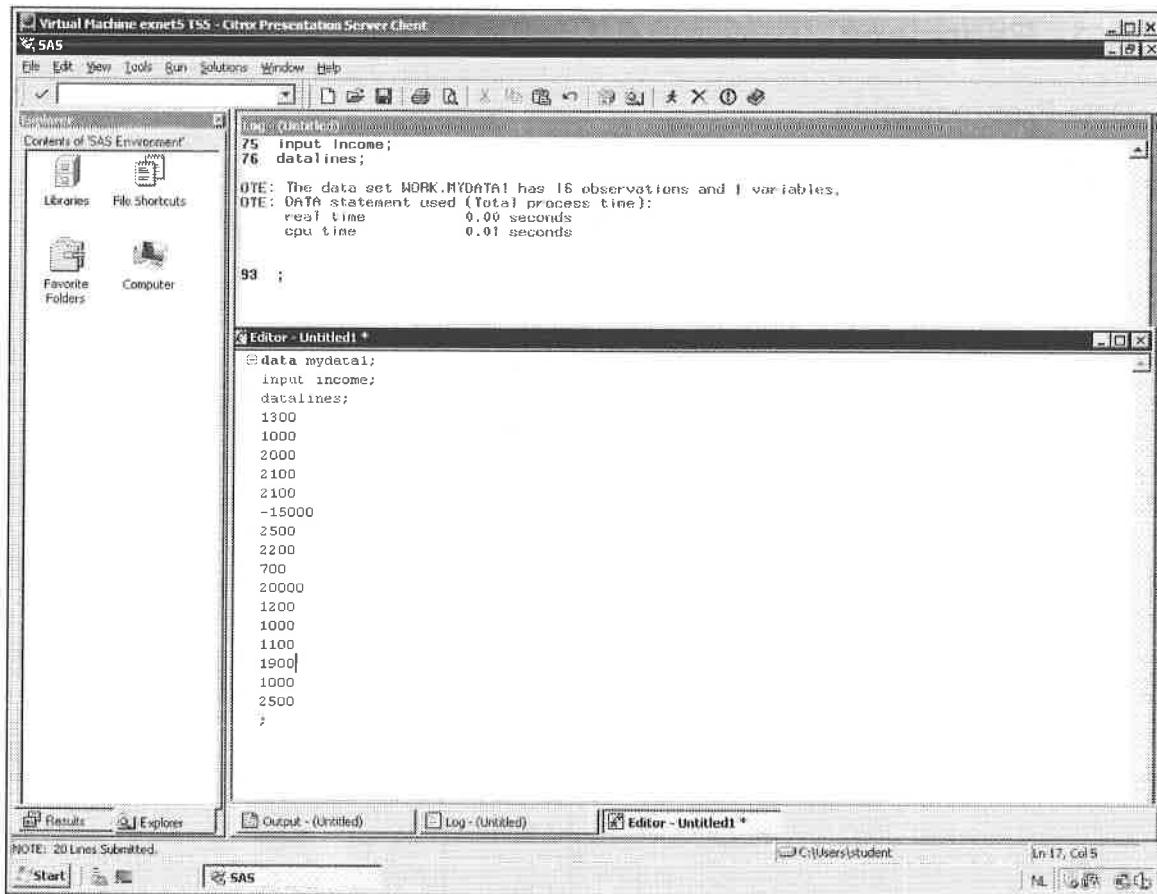
A.1 Exercises

Data Preprocessing

1. Outliers – Univariate

- a. Create the following SAS data set and give it the name **mydata1**:
(See the course handouts for the SAS code.)

Income
1300
1000
2000
2100
-15000
2500
2200
700
20000
1200
1000
1100
1900
1000
2500



- b. Check for univariate outliers by using a histogram and box plot in SAS/INSIGHT. Select **Solutions** from the menu above, and then select **Analysis** \Rightarrow **Interactive Data Analysis**. Select the Work library and the **mydata1** data set that you just created. In the new menu above, select **Analyze**, **Histogram/Bar Chart**, and create a histogram. After you are finished, go back and select **Analyze**, **Box Plot/Mosaic Plot**, and create a box plot.
- c. Calculate the z-score and use PROC STANDARD.

```

proc standard data=mydata1 mean=0 std=1
              out=stdmydata1;
run;

```

- d. Identify outliers that have $|z| > 2.5$ as follows:

```
data outliers;
set stdmydata1;
where abs(Income) > 2.5;
run;
```

The screenshot shows the SAS software interface running on a Windows operating system. The title bar reads "Virtual Machine exnet5 TSS - Citrix Presentation Server Client". The menu bar includes File, Edit, View, Tools, Run, Solutions, Window, and Help. The left pane is an "Explorer" window titled "Contents of 'Work'" showing various datasets: Mydata, Mydata1, Mydata2, Mydata3, Mydata4, Outliers, Stdmydata, Stdmydata1, Stdmydata3, and Stdmydata4. The right pane has two windows: "Log - (Untitled)" and "Editor - Untitled1 *". The Log window displays the output of the run command and the results of the WHERE clause. The Editor window shows the SAS code. The bottom status bar indicates "NOTE: 7 Lines Submitted.", the current directory "C:\Users\student", and the cursor position "ln 22, Col 56".

NOTE: There were 2 observations read from the data set WORK.STDMYDATA1,
WHERE ABS(INCOME)>2.5;
NOTE: The data set WORK.UTLTI1.FRR has 2 observations and 1 variables.

```
data mydata1;
input income;
datalines;
1300
1000
2000
2100
2100
-15000
2500
2200
700
20000
1200
1000
1100
1900
1000
2500
;

proc standard data=mydata1 mean=0 std=1 out=stdmydata1;
run;

data outliers;
set stdmydata1;
where abs(income) > 2.5;
run;
```

2. Outliers - Multivariate

- a. Create the following SAS data set:
(See the course handouts for the SAS code.)

Income	Savings Amount
1300	400
1000	350
2000	800
2100	200
1700	600
2500	1000
2200	900
700	1000
1500	500

The screenshot shows the SAS software interface. On the left, there is a sidebar with icons for Libraries, File Shortcuts, Favorite Folders, and Computer. The main window has three panes: 'Log - (Untitled)' at the top, 'Editor - Untitled1 *' in the center, and 'Results' at the bottom. The 'Log' pane displays the SAS log output:

```
124 data mydata2;
125 input income savings;
126 datalines;
```

NOTE: The data set WORK.MYDATA2 has 10 observations and 2 variables.
NOTE: DATA statement used (Total process time):
 real time 0.00 seconds
 cpu time 0.00 seconds

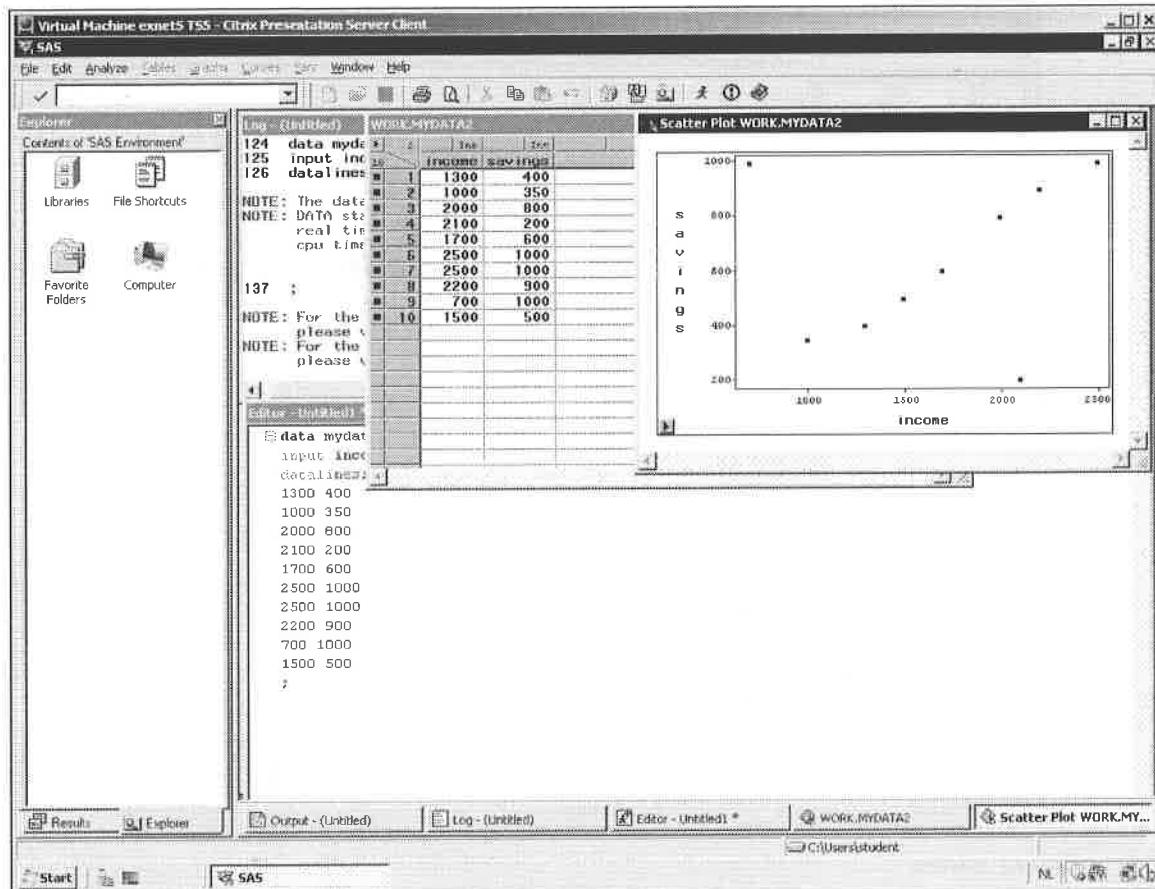
137 ;

The 'Editor' pane contains the SAS code and data:

```
data mydata2;
  input income savings;
  datalines;
1300 400
1000 350
2000 800
2100 200
1700 600
2500 1000
2500 1000
2200 900
700 1000
1500 500
;
```

The 'Results' pane at the bottom shows the note: 'NOTE: 14 Lines Submitted.'

- b. Check for multivariate outliers by using a scatter plot in SAS/INSIGHT. Select **Solutions** from the menu and select **Analysis, Interactive Data Analysis**. Select the **Work** library and then the **mydata2** data set that you created. In the new menu, select **Analyze, Scatter Plot**, and create a scatter plot as follows:

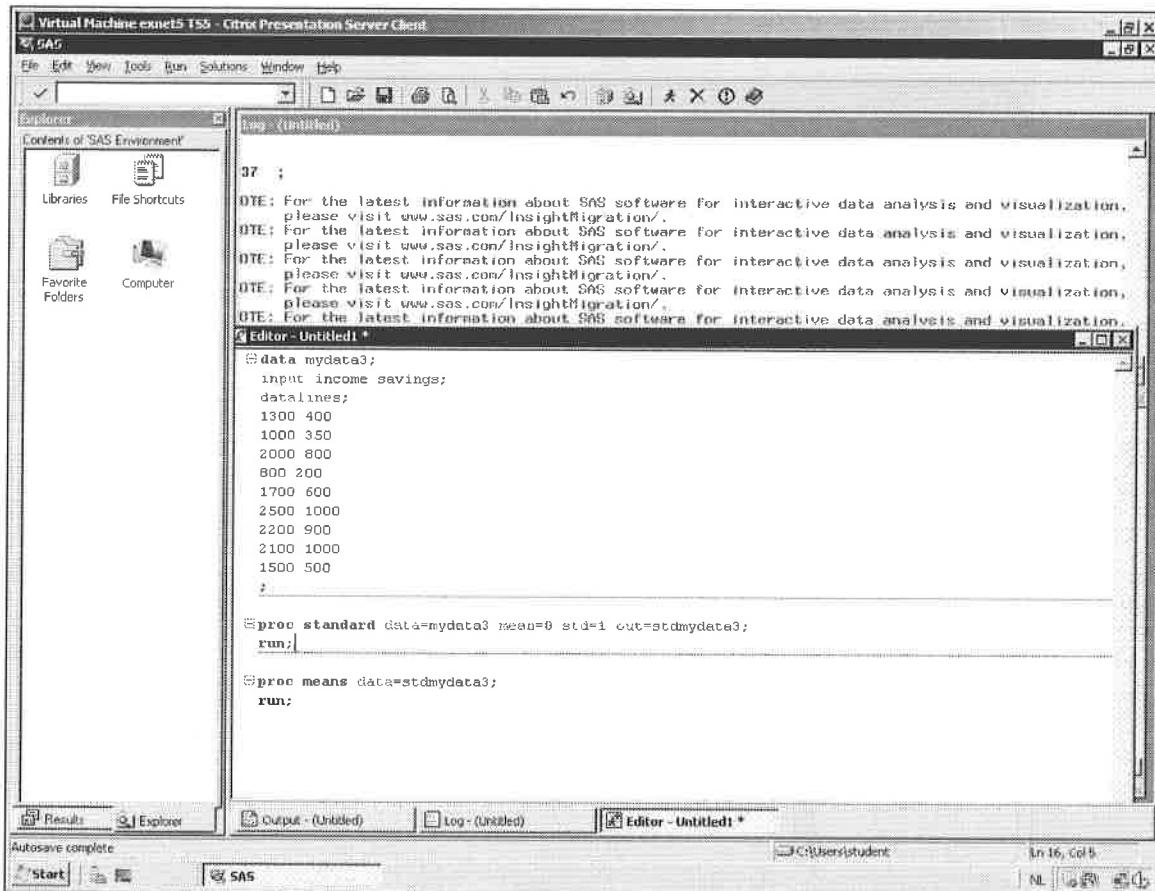


3. Standardization

- a. Statistically standardize the following credit scoring data using PROC STANDARD:

Income	Savings Amount
1300	400
1000	350
2000	800
800	200
1700	600
2500	1000
2200	900
2100	1000
1500	500

- b. Check that the mean and standard deviation of the standardized data are equal to 0 and 1, respectively. Use PROC MEANS.



The screenshot shows the SAS 9.13 software interface. The top menu bar includes File, Edit, View, Tools, Run, Solutions, Window, and Help. The left sidebar has sections for Libraries, File Shortcuts, and Favorite Folders. The main window contains an 'Editor - Untitled1' tab where the following SAS code is written:

```

37 ;
DTE: For the latest information about SAS software for interactive data analysis and visualization,
please visit www.sas.com/insightmigration/.
DTE: For the latest information about SAS software for interactive data analysis and visualization,
please visit www.sas.com/insightmigration/.
DTE: For the latest information about SAS software for interactive data analysis and visualization,
please visit www.sas.com/insightmigration/.
DTE: For the latest information about SAS software for interactive data analysis and visualization,
please visit www.sas.com/insightmigration/.
DTE: For the latest information about SAS software for interactive data analysis and visualization,
please visit www.sas.com/insightmigration/.
DTE: For the latest information about SAS software for interactive data analysis and visualization,
please visit www.sas.com/insightmigration/.
Editor - Untitled1*
data mydata3;
input income savings;
datalines;
1300 400
1000 350
2000 800
800 200
1700 600
2500 1000
2200 900
2100 1000
1500 500
;

proc standard data=mydata3 mean=0 std=1 out=stdmydata3;
run;

proc means data=stdmydata3;
run;

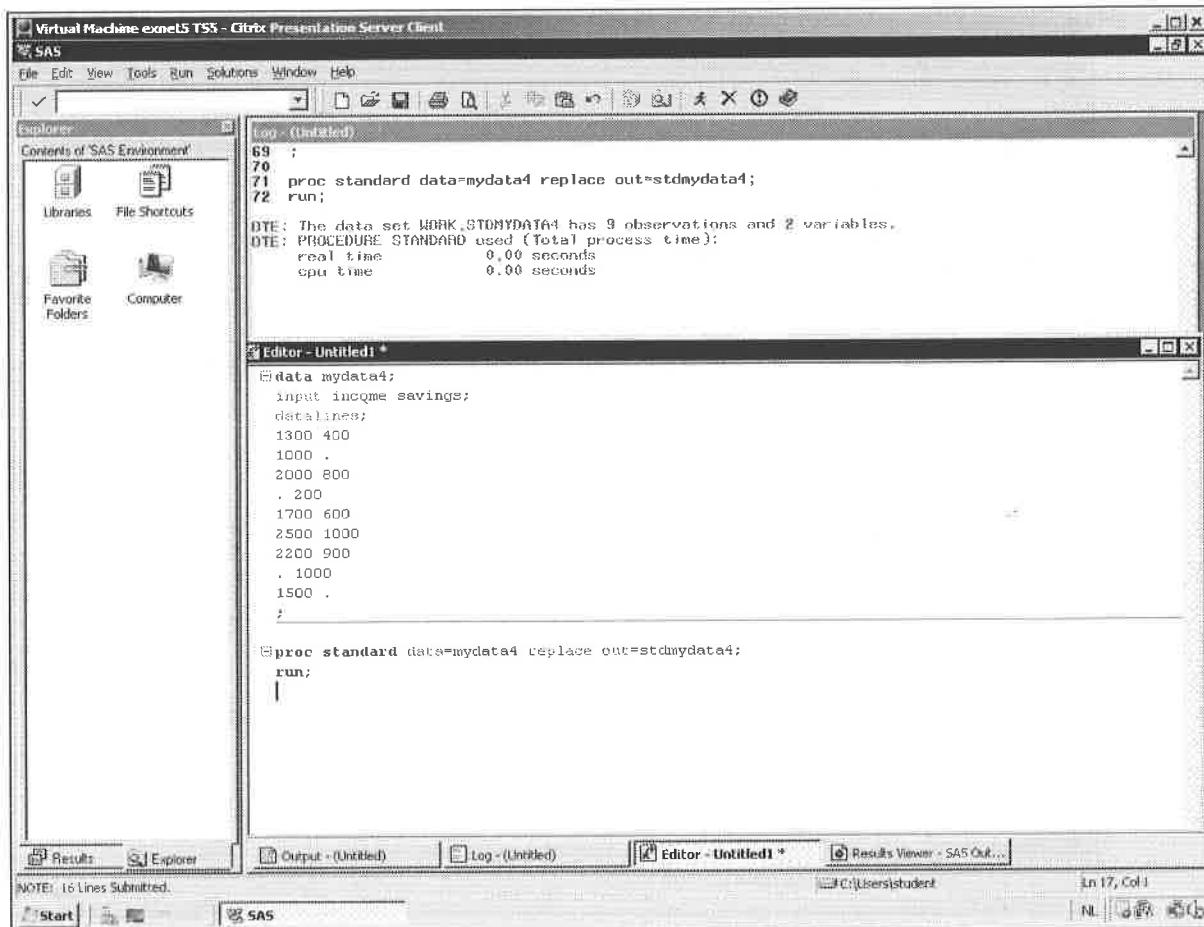
```

The bottom status bar indicates the current location is C:\Users\student\ at line 16, column 5. The bottom tabs show Results, Explorer, Output - Untitled, Log - Untitled, and Editor - Untitled1*. The bottom toolbar includes Start, Stop, and SAS buttons.

4. Missing Values

Income	Savings Amount
1300	400
1000	.
2000	800
.	200
1700	600
2500	1000
2200	900
.	1000
1500	.

Replace the missing values with the mean using PROC STANDARD with the REPLACE option.
 (See the course handouts for the SAS code.)



The screenshot shows the SAS software interface. The top menu bar includes File, Edit, View, Tools, Run, Solutions, Window, and Help. The left sidebar has sections for Explorer (Libraries, File Shortcuts, Favorite Folders, Computer), Output (Output - Untitled), Log (Log - Untitled), and Editor (Editor - Untitled1). The Editor window contains the following SAS code:

```

69 ;
70 proc standard data=mydata4 replace out=stdmydata4;
71 run;

NOTE: The data set WORK/stdmydata4 has 9 observations and 2 variables.
NOTE: PROCEDURE STANDARD used (Total process time):
      real time           0.00 seconds
      cpu time            0.00 seconds
      
```

The Editor window also contains a data step:

```

data mydata4;
  input income savings;
  datalines;
  1300 400
  1000 .
  2000 800
  . 200
  1700 600
  2500 1000
  2200 900
  . 1000
  1500 .
  ;

```

```

proc standard data=mydata4 replace out=stdmydata4;
run;
  
```

The bottom status bar shows "NOTE: 16 Lines Submitted.", the current working directory "C:\Users\student", and the cursor position "Ln 17, Col 1".

5. Coarse Classification

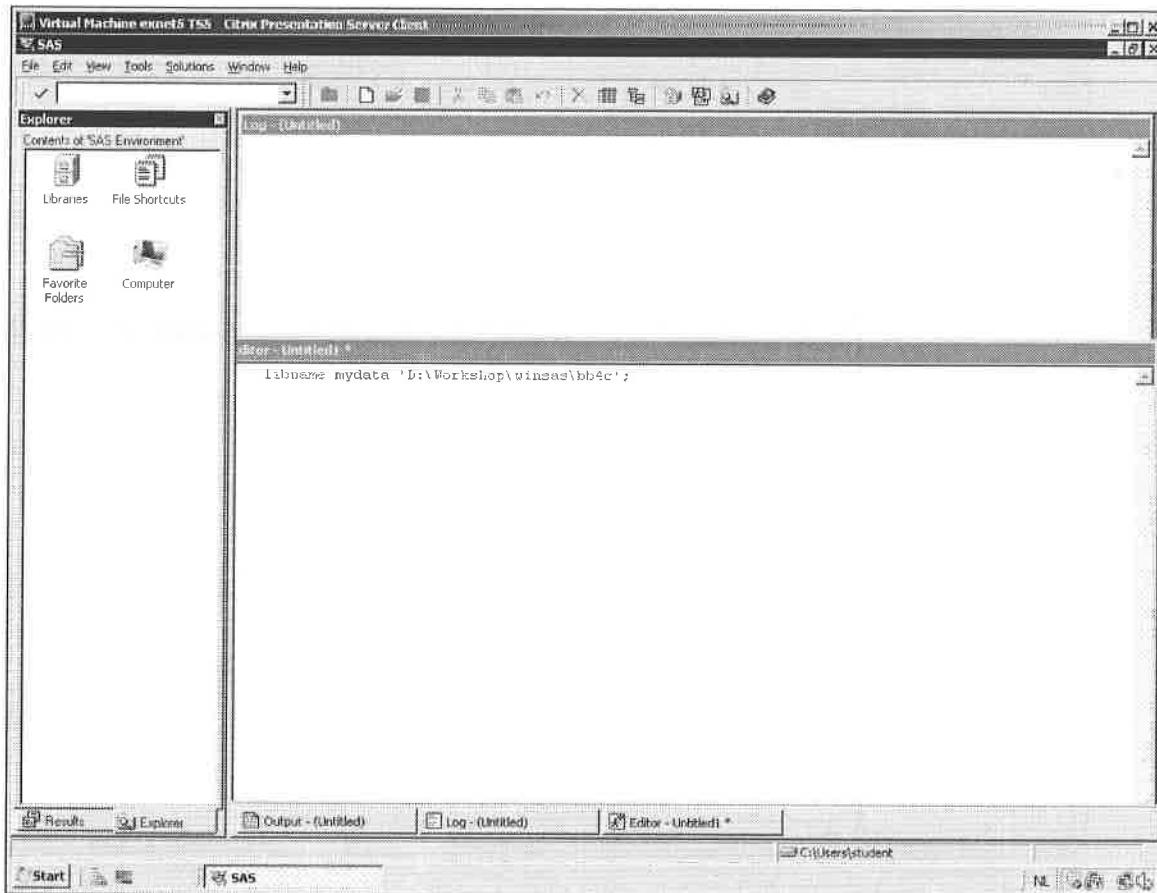
- a. Suppose you want to coarse classify the attribute number of children, which has the following distribution:

Attribute	No children	1-2 children	3+ children
Number of goods	1500	2200	300
Number of bads	500	300	200

- b. You consider splitting the variable into two groups: either no children versus children, or two or fewer children versus three or more children. Find the best split using PROC FREQ in SAS/STAT. (See the course handouts about coarse classification for the SAS code.)

6. Preprocessing an Application Scoring Data Set

- a. Create a SAS library **mydata** as follows:



- b. Run the LIBNAME statement. The **mydata** library contains all the data sets that will be analyzed in what follows. One of them is the **applicants** data set. This is a publicly available credit scoring data set containing application characteristics (for example, age, amount of loan, and purpose of loan) and a target variable indicating whether the applicant was a good payer or a bad payer. (You can find the data dictionary of this data set in Appendix B.)
- c. Explore the **applicants** data set by double-clicking **Libraries** in the Explorer panel (left part of the display), double-clicking the **mydata** library, and double-clicking the **applicants** data set.
- d. Start SAS/INSIGHT by selecting **Solutions** from the menu, and then select **Analysis, Interactive Data Analysis**. Select the **mydata** library and the **applicants** data set.
- e. In the menu above, select **Analyze** and create the following:
- 1) Histogram of the **age** variable
 - 2) Histogram of the **purpose** variable
 - 3) Box plot of the **amount** variable
 - 4) Box plot for the **age** variable for the goods and bads separately (Use the **good_bad** variable as a group criterion.)
 - 5) Scatter plot of the **amount** versus the **duration** variable

- f. Try some of the other options provided by SAS/INSIGHT (for example, **Analyze, Distribution**, and **Analyze, Fit** to estimate a linear regression model relating, for example, to amount to duration).

7. The Basel II Capital Requirement Formulas

- a. Create a SAS macro as follows:

```
%macro BaselCap(PD, LGD, EAD, corr);
  temp1=((1/(1-&corr))**0.5)*probit(&PD) +
    ((&corr/(1-&corr))**0.5)*probit(0.999));
  temp2=CDF('Normal',temp1);
  Cap=&EAD*&LGD*(temp2-&PD);
%mend BaselCap;
```

This macro will compute the Basel II capital requirements for a given PD, LGD, and correlation factor.

- b. Create the following SAS data sets computing the Basel II capital requirements for residential mortgages and qualifying revolving exposures for low PDs:

```
data resmortgage;
  do i=1 to 200;
  PD=i/1000;
  %BaselCap(PD,LGD=0.50,EAD=1,corr=0.15);
  output;
end;

data QRE;
  do i=1 to 200;
  PD=i/1000;
  %BaselCap(PD,LGD=0.50,EAD=1,corr=0.04);
  output;
end;
```

- c. Plot the Basel II capital requirements for residential mortgages and qualifying revolving exposures as follows:

```
proc gplot data=resmortgage;
  plot Cap*PD;
run;

proc gplot data=QRE;
  plot Cap*PD;
run;
```

- d. Consider a credit card portfolio with the following true characteristics: PD=0.05; LGD=0.20 and EAD=\$10,000. (Remember that credit cards are revolving credits for which corr=0.04.)

- 1) Compute the capital requirement assuming the following estimates:

Scenario	PD	LGD	EAD	Capital Requirement
Everything correct	0.05	0.20	10,000	
10% overestimate PD	0.055	0.20	10,000	
10% overestimate LGD	0.050	0.22	10,000	
10% overestimate EAD	0.050	0.20	11,000	

```
data example;
  %BaselCap(PD=0.05,LGD=0.20,EAD=10000,corr=0.04); output;
  %BaselCap(PD=0.055,LGD=0.20,EAD=10000,corr=0.04); output;
  %BaselCap(PD=0.05,LGD=0.22,EAD=10000,corr=0.04); output;
  %BaselCap(PD=0.05,LGD=0.20,EAD=11000,corr=0.04); output;
run;
```

- 2) Which scenario(s) has the biggest impact on the capital requirement? Why?

The screenshot shows the SAS Editor window with the following code:

```

Virtual Machine exact5 TSS - Citrix Presentation Server Client
SAS - [Editor - Untitled] *
File Edit View Tools Run Solutions Window Help
Contents of SAS Environment
Libraries File Shortcuts
Favorite Folders Computer
macro BaselCap (PD, LGD, EAD, corr);
temp1=((1/(1-corr))**0.5)*probit(sPD) + ((scorr/(1-corr))**0.5)*probit (-0.999);
temp2=CDF('Normal', temp1);
Cap=&EAD*&LGD*(temp2-&PD);
%end BaselCap;

data remmortgage;
do i=1 to 200;
PD=i/1000;
%BaselCap(PD, LGD=0.50, EAD=i, corr=0.04);
output;
end;
run;

data QRE;
do i=1 to 200;
PD=i/1000;
%BaselCap(PD, LGD=0.50, EAD=i, corr=0.04);
output;
end;
run;

proc gplot data=remmortgage;
plot Cap*PD;
run;

data example;
%BaselCap(PD=0.05, LGD=0.20, EAD=10000, corr=0.04); output;
%BaselCap(PD=0.055, LGD=0.20, EAD=10000, corr=0.04); output;
%BaselCap(PD=0.05, LGD=0.22, EAD=10000, corr=0.04); output;
%BaselCap(PD=0.05, LGD=0.20, EAD=11000, corr=0.04); output;
run;

```

The bottom status bar indicates "NOTE: 6 Lines Submitted." and the bottom right corner shows "C:\Users\bbaesens" and "In 34, Col 1".

Classification Techniques

8. Logistic Regression

- a. Estimate a logistic regression classifier using PROC LOGISTIC for the **applicants** data set. (See Appendix B.) The dependent variable is the **good_bad** variable. The independent variables are **checking** and **savings** (both nominal), and **duration**, **amount**, and **age** (all three continuous). Include the PARAM=GLM option in the CLASS statement to indicate the type of 0/1 dummy encoding to be used for the class variables. Ask for a confusion matrix using the CTABLE option.

Your SAS code should look as follows:

```
libname mydata 'd:\workshop\winsas\bb4c';
proc logistic data=mydata.applicants;
  class checking savings / param=glm;
  model good_bad=checking duration amount savings age /ctable;
run;
```

What is the classification accuracy assuming a cut-off of 0.5? Check which are the important variables using the *p*-values. (Remember that low *p*-values indicate important variables.)

- b. Now estimate a logistic regression classifier using a PROBIT and CLOGLOG link function. (Use LINK=PROBIT and LINK=CLOGLOG at the end of the MODEL statement.)

```
libname mydata 'd:\workshop\winsas\bb4c';
proc logistic data=mydata.applicants;
  class checking savings / param=glm;
  model good_bad=checking duration amount savings age
    /link=probit ctable;
run;
```

Investigate the impact on the classification performance assuming a cut-off of 0.5.

- 1) Which link function gives the best performance?
- 2) Which are the most important variables?
- 3) Are these the same across all estimated models?

The screenshot shows the SAS Enterprise Miner interface running on a Citrix Presentation Server Client. The main window title is "Virtual Machine ex01e5 TSS - Citrix Presentation Server Client". The menu bar includes File, Edit, View, Tools, Run, Solutions, Window, and Help. The interface consists of several panes:

- Results** pane: Shows the output of a PROC LOGISTIC run. The code is:

```
proc logistic data=mydata.applicants;
class checking savings/param=glm;
model good_bad=checking duration amount savings age/ctable;
run;
```

Notes from the log include:
 - NOTE: Writing HTML Body file: sashtml.htm
 - NOTE: PROC LOGISTIC is modeling the probability that good_bad='bad'. One way to change this to model the probability that good_bad='good' is to specify the response variable option EVENT='good'.
 - NOTE: Convergence criterion (GCONV=1E-8) satisfied.
 - NOTE: There were 1000 observations read from the data set MYDATA.APPLICATIONS.
 - NOTE: PROCEDURE LOGISTIC used (Total process time):
 - real time 0.49 seconds
 - cpu time 0.46 seconds
- Log - (Untitled)** pane: Displays the log output of the PROC LOGISTIC run, identical to the Results pane.
- Editor - Untitled1** pane: Shows the original SAS code:

```
proc logistic data=mydata.applicants;
class checking savings/param=glm;
model good_bad=checking duration amount savings age/ctable;
run;
```
- Output - (Untitled)** pane: Shows the output of the PROC LOGISTIC run, identical to the Results pane.
- File Explorer** pane: Shows the file structure at "C:\Users\student".
- SAS** status bar: Shows "Ln 6, Col 5".

9. Linear Programming

- a. Consider the following data set in SAS:

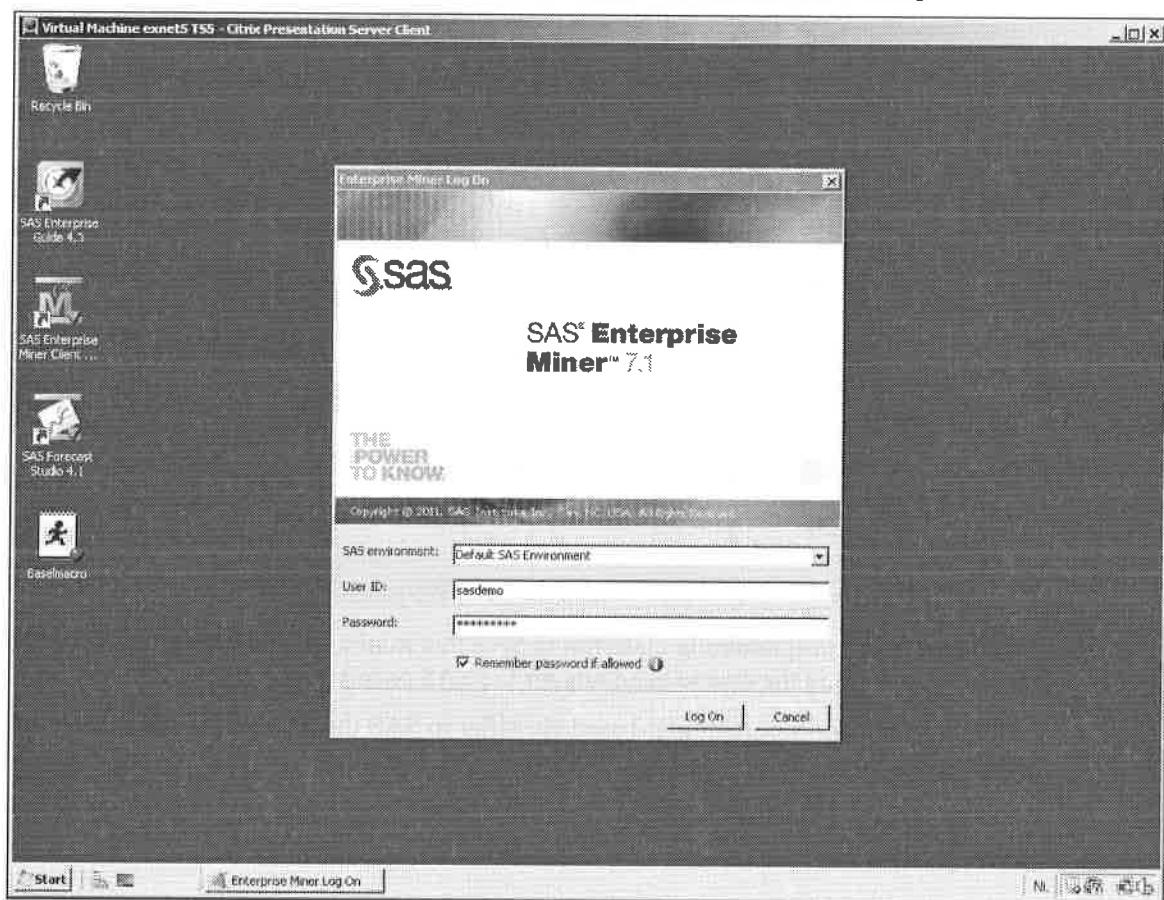
Age	Years at address	Phone	Good/Bad
24	2	No	Bad
35	6	Yes	Bad
40	4	Yes	Bad
30	10	No	Good
28	1	No	Bad
50	5	Yes	Bad
45	15	Yes	Good
60	10	No	Good
20	1	No	Bad
25	5	Yes	Good

- b. Estimate a linear programming classifier in SAS that minimizes the maximum deviation.
(See the example in the course handouts for the SAS code.)
- c. Estimate an integer-programming-based classifier in SAS that minimizes the number of misclassifications.
- d. Suppose that previous experience indicates that age is more important than either of the other two characteristics. How would you ensure that and what is the impact on the results?

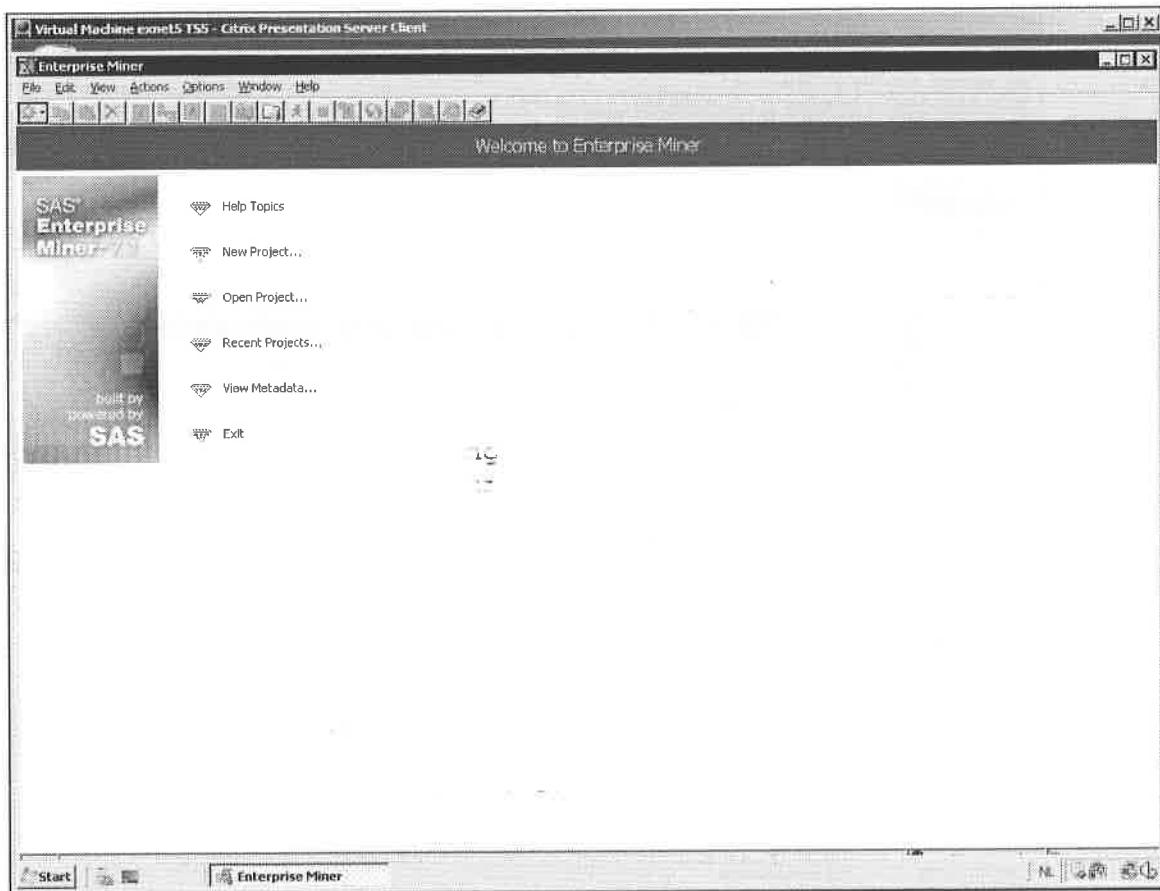
Using SAS Enterprise Miner

10. Exploring the Tool

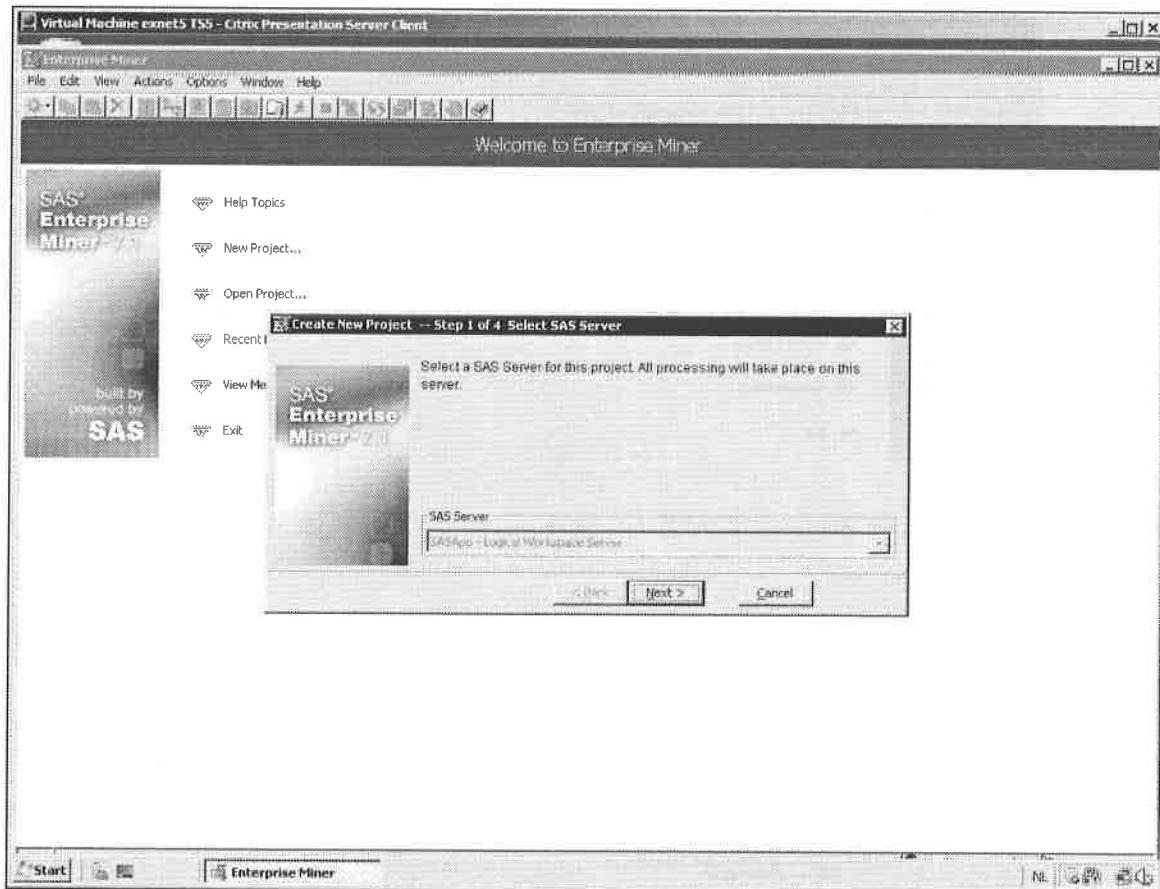
- Start SAS Enterprise Miner 7.1 by double-clicking the shortcut on the desktop.



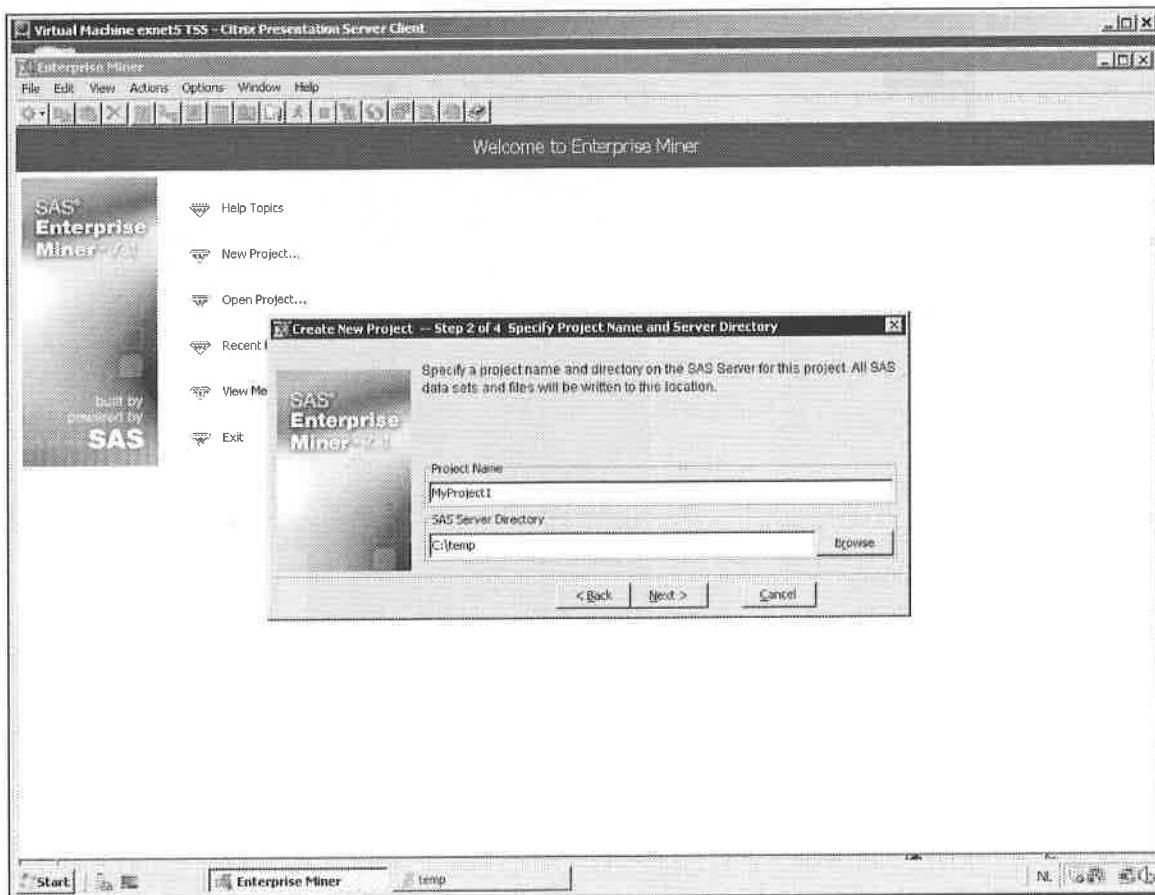
- b. Click **Log On** with the provided user name and password. You see the following display:



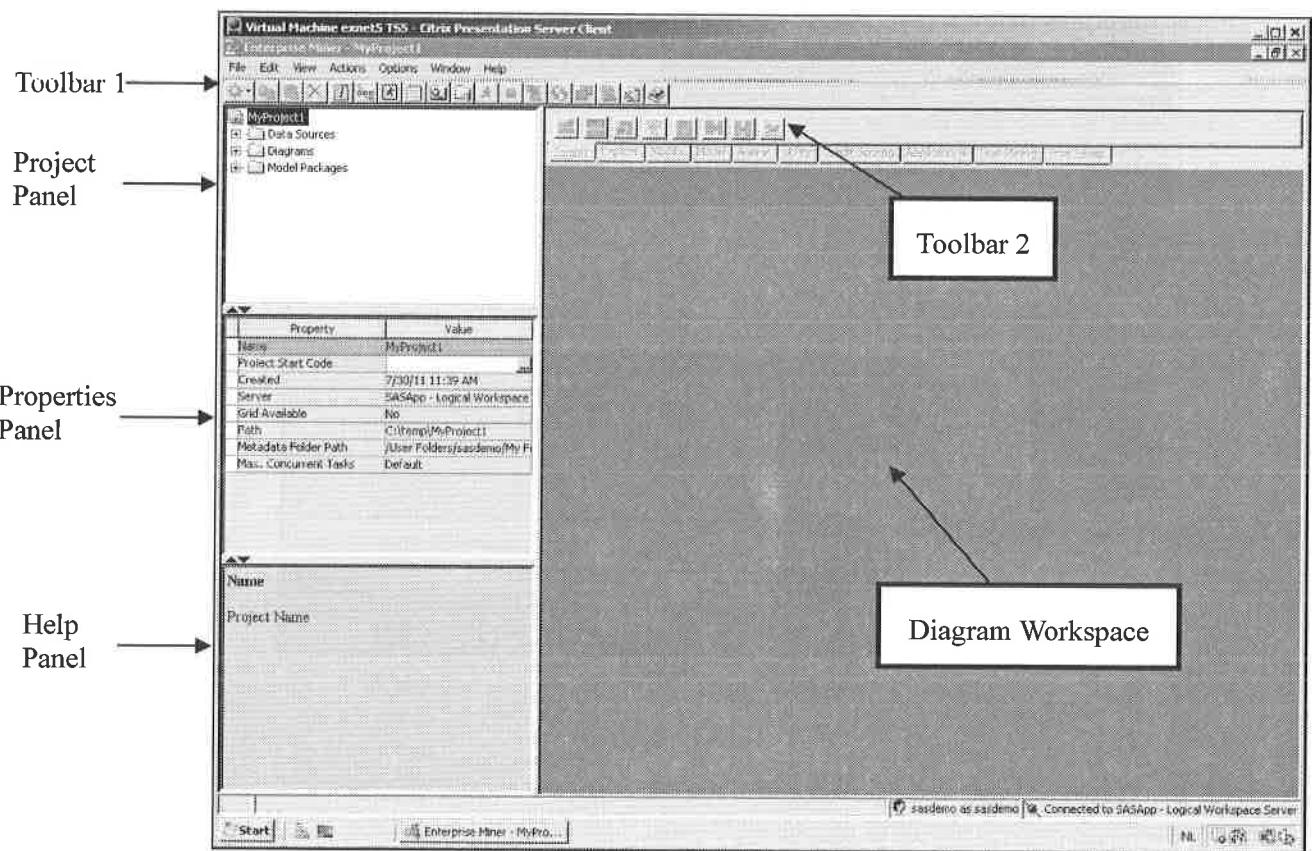
- c. Select New Project and create a new project as follows:



- d. Click **Next**. Specify the name of the project as **MyProject1** and the SAS Server Directory as **C:\temp** as follows:



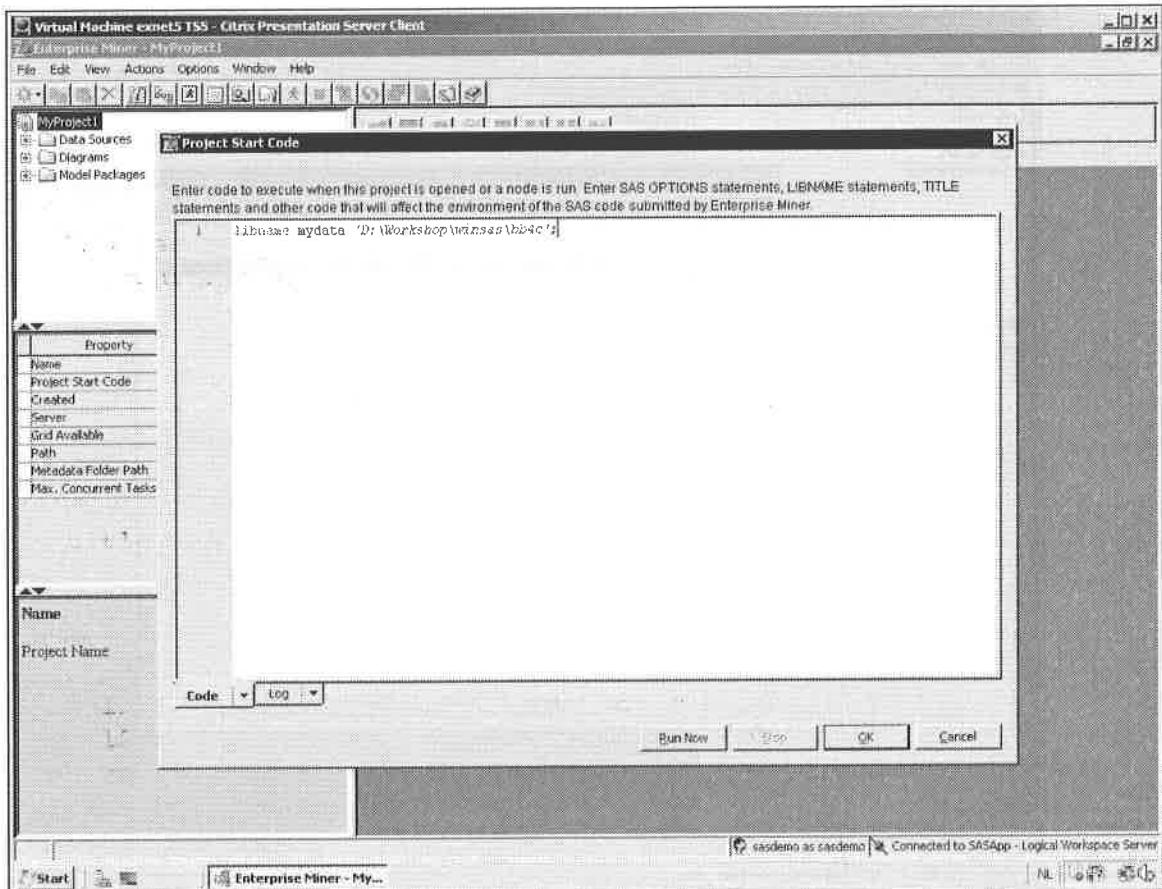
- e. Click **Next**, and finish the wizard. The following display appears:



The interface is divided into five components:

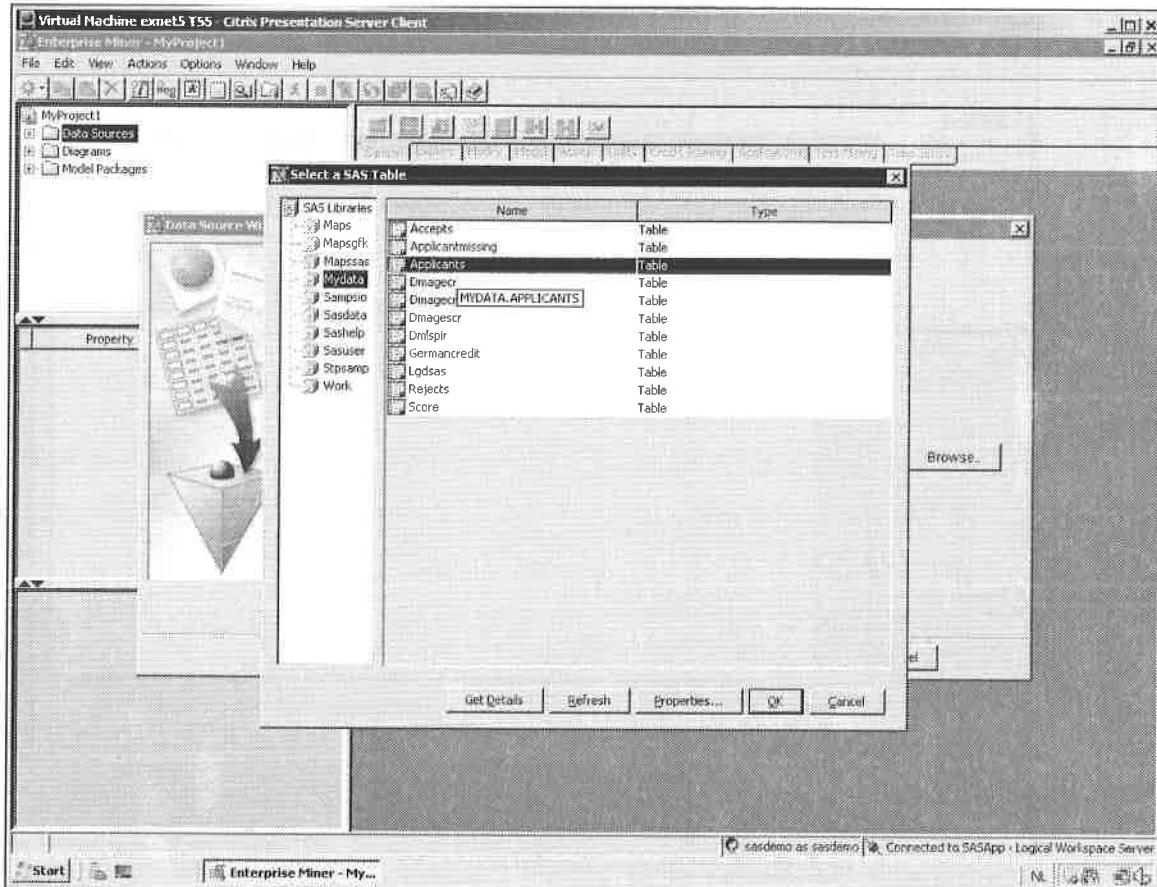
- **Toolbar 1** – This toolbar provides a set of common utilities to assist in building your projects.
- **Toolbar 2** – This toolbar is a graphical set of node icons and tools that you use to build process flow diagrams in the Diagram Workspace. To display the text name of any node or tool icon, position your mouse pointer over the icon.
- **Project panel** – Use the Project panel to manage and view data sources, diagrams, model packages, and list users.
- **Properties panel** – Use the Properties panel to view and edit the settings of any object that you select (including data sources, diagrams, nodes, results and users).
- **Diagram Workspace** – Use the diagram workspace to build, edit, run, and save process flow diagrams. This is where you graphically build, order, and sequence the nodes that you use to mine your data and generate reports.
- **Help panel** – The Help panel displays a short description of the property that you select in the Properties panel. Extended help can be found in the Help Topics selection from the Help main menu.

- f. Create a library to the physical directory on disk containing the data sets. In the Properties panel, click on the button next to the **Project Start Code Property**, and then type the code shown in the following display:



- g. Click **Run Now** \Rightarrow **OK**. The **mydata** library is successfully created.
- h. Right-click on the **Data Sources** folder in the Project panel (or select **File** \Rightarrow **New** \Rightarrow **Data Source**) to open the Data Source Wizard. Create a new data source for the **applicants** data set, which is located in the **mydata** library. (This is a publicly available credit scoring data set containing application characteristics (for example, age, amount of loan, and purpose of loan) and a target variable indicating whether the applicant was a good payer or a bad payer. You can find the data dictionary of this data set in Appendix B.)

- i. In step 1 of the wizard, click **Next**. In step 2 of the wizard, click **Browse** and select the **mydata** library that you created. If needed, click the **Refresh** button at the bottom of the display to see the various data sets. You see the following:

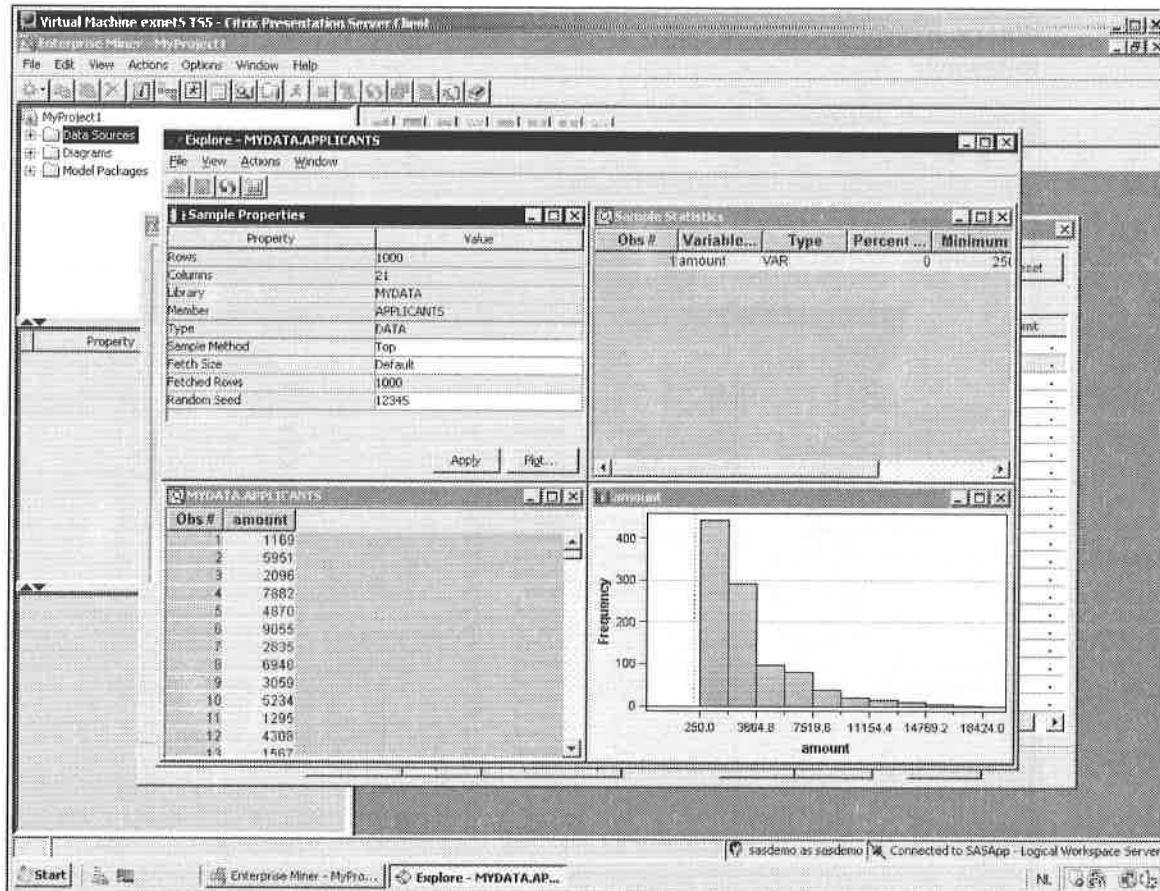


- j. Click **OK** \Rightarrow **Next**. In step 3 of the wizard, click **Next**. In step 4 of the wizard, click **Next**.

- k. Set the level of the **checking** and **savings** variables to **ordinal**. Set the level of the **coapp**, **employed**, **foreign**, **history**, **housing**, **job**, **marital**, **other**, **property**, **purpose**, **resident**, and **telephon** variables to **nominal**. Set the role of the **good_bad** variable to **target** and its level to **binary**. Accept the Interval (=continuous) level suggested for the other variables. You have the following:

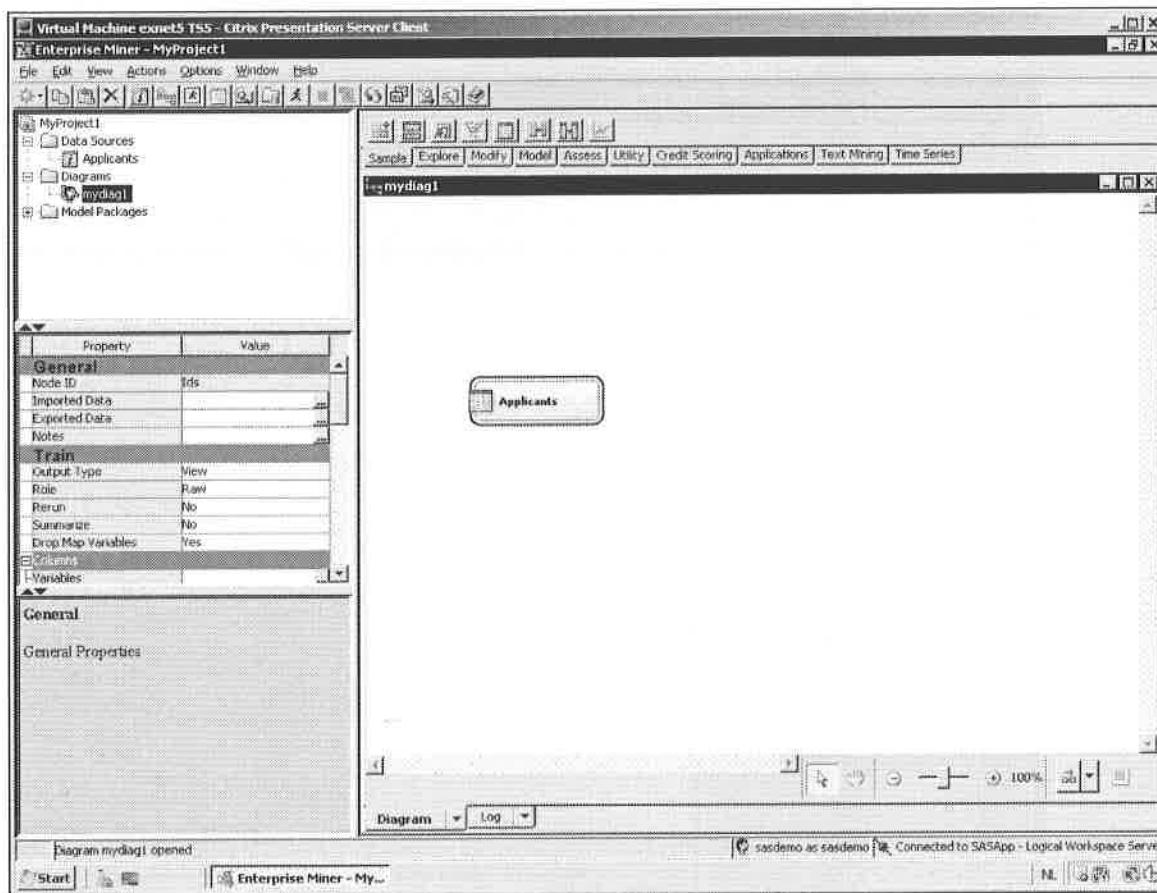
Name	Role	Level	Report	Order	Drop	Lower Limit	Upper Limit
age	Input	Interval	No	No		.	.
amount	Input	Interval	No	No		.	.
checking	Input	Ordinal	No	No		.	.
coapp	Input	Nominal	No	No		.	.
depends	Input	Interval	No	No		.	.
duration	Input	Interval	No	No		.	.
employed	Input	Nominal	No	No		.	.
existing	Input	Interval	No	No		.	.
foreign	Input	Nominal	No	No		.	.
good_bad	Target	Binary	No	No		.	.
history	Input	Nominal	No	No		.	.
housing	Input	Nominal	No	No		.	.
installip	Input	Interval	No	No		.	.
job	Input	Nominal	No	No		.	.
marital	Input	Nominal	No	No		.	.
other	Input	Nominal	No	No		.	.
property	Input	Nominal	No	No		.	.
purpose	Input	Nominal	No	No		.	.
resident	Input	Nominal	No	No		.	.
savings	Input	Ordinal	No	No		.	.
telephon	Input	Nominal	No	No		.	.

- I. Click on the **amount** variable and explore its distribution by clicking  at the bottom of the display.

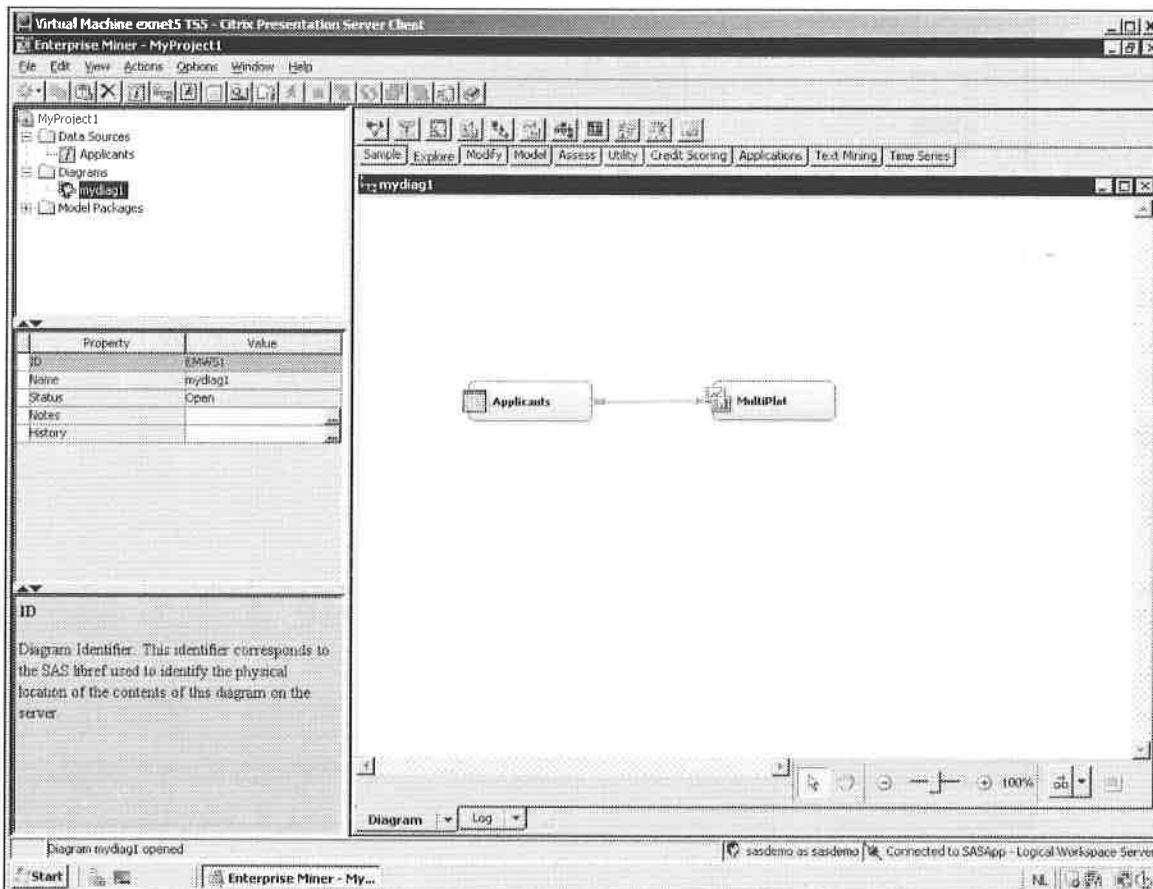


- m. Close this window, and then click **Next** in steps 5, 6, 7, and 8, and **Finish** in step 9. The data was successfully entered and is ready to be analyzed.

- n. In the project panel, right-click **Diagrams** and create a new diagram named **mydiag1**. Drag and drop the **applicants** data set from the project to the diagram workspace. You have the following:



- o.** Go to the Explore tab of toolbar 2. Add a Multiplot node to the diagram workspace and connect it to the **applicants** data set. Right-click the **Multiplot** node and select **Run**. After the run is finished, right-click the node again, and select **Results**. Inspect the graphs and output generated by this node. Your workspace should look as follows:



- p.** Select the **Sample** tab and add a Sample node to the diagram workspace. Connect it to the **applicants** data set. Set the Sample method in the Properties panel to stratify to create a stratified sample based on the target variable **good_bad**. Run the node and inspect the generated sample by clicking the button next to the **Exported Data** property in the Properties panel.
- q.** From the Sample tab, add a Filter node to the diagram workspace and connect it to the Applicants node. Set its properties to eliminate class variable values that occur < 2 times (the Minimum Frequency Cutoff property=2) for class variables with < 20 different values (the Maximum Number of levels Cutoff property=20). This means that if a class variable has fewer than 20 different values (for example, marital status), then all observations having class values that occur fewer than two times are removed.
- 1) Also eliminate interval variable values that are more than 2.5 standard deviations away from the mean. (Click the button next to the **Tuning Parameters** property.)
 - 2) Run the Filter node and inspect the results.
 - 3) Inspect the limits for the interval variables and check whether any class values were excluded.
 - 4) Check how many total observations were removed.

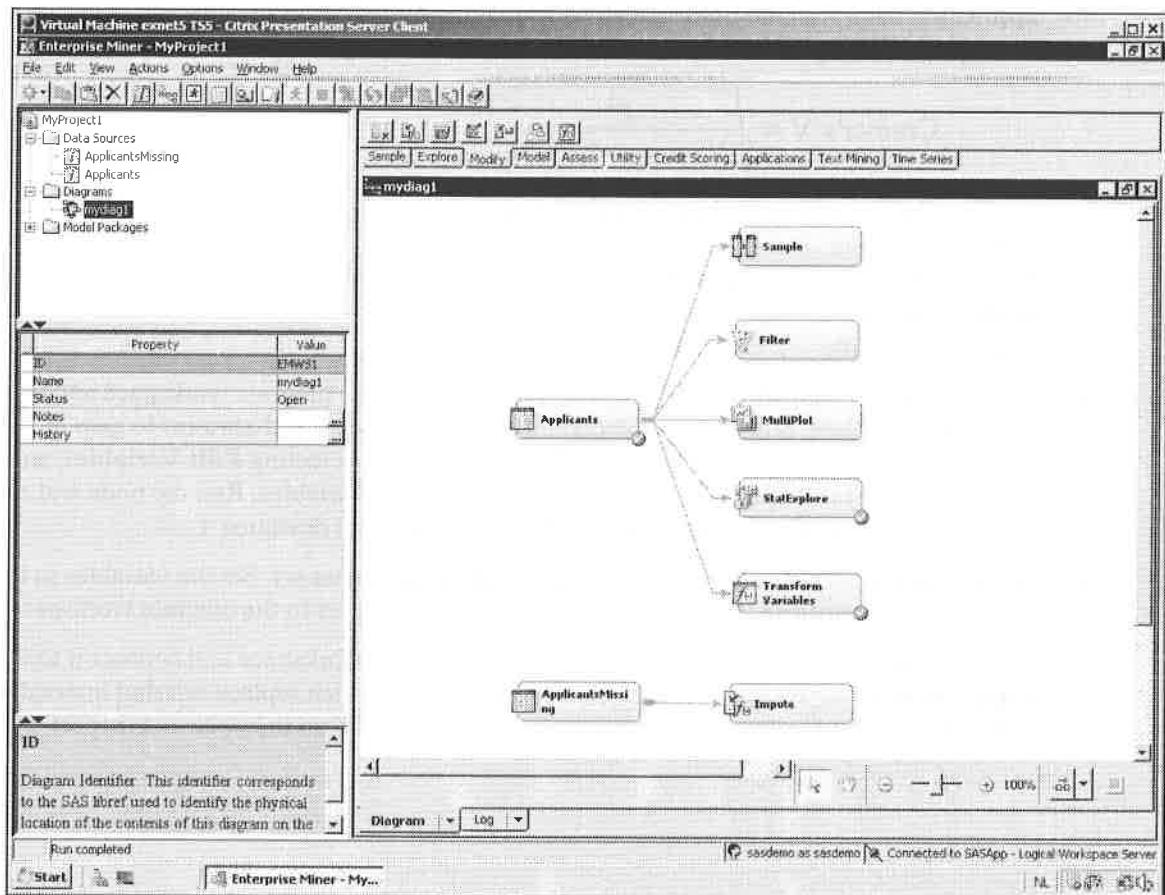
- r. From the Explore tab, add a StatExplore node to the diagram workspace and connect it to the **applicants** data. Run the StatExplore node and inspect the results. A chi-square plot is presented and reports Cramer's V statistic. The Cramer's V statistic is computed as follows:

$$\text{Cramer's } V = \sqrt{\frac{\chi^2}{N \min(r-1, c-1)}}$$

N is the number of observations, r is the number of rows, and c is the number of columns of the contingency table. The statistic varies between 1 (perfectly predictive input) and 0 (useless input). Inspect the other output generated by this node. Notice that, for the checking account variable, the mode is different for the goods and bads separately, which clearly illustrates that this is an important variable.

- s. From the Modify tab, add a Transform Variables node to the diagram workspace and connect it to the **applicants** data. Standardize the continuous variables **age** and **amount** to zero mean and unit standard deviation by right-clicking the **Transform** node, selecting **Edit Variables**, and then set the method to **Standardize** for both the **age** and **amount** variables. Run the node and check whether the mean of the new variables is 0 and the standard deviation 1.
- t. Create a new Input data source for the **applicantsmissing** data set. Set the variables in the same way as for the **applicants** data set. Drag and drop the data set to the diagram workspace.
- u. From the Modify tab, add an Impute node to the diagram workspace and connect it to the **applicantsmissing** data set. Accept the default settings, which replace missing interval variables with the **mean** and **missing** class variables with the mode. Run the node and inspect the results.

Your workspace should now look as follows:

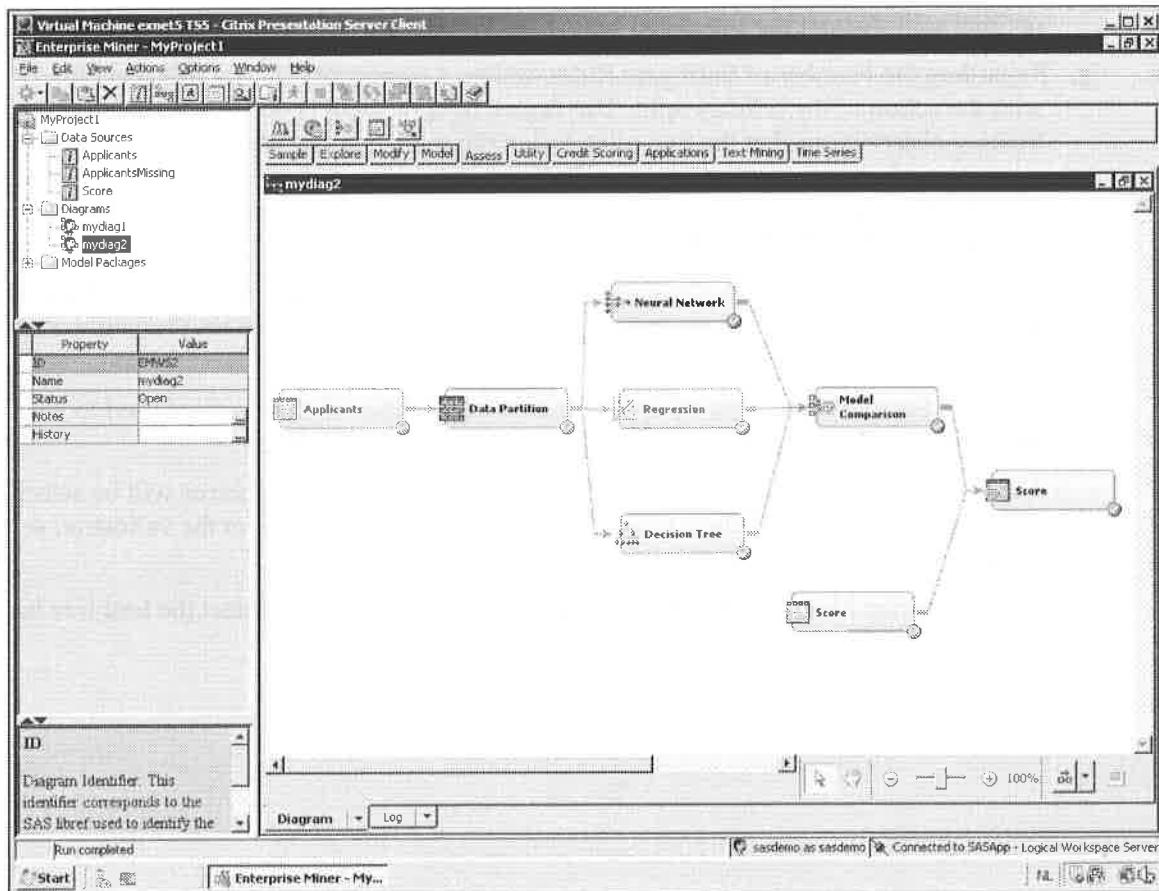


11. Developing a First Credit Scoring Model

- Create a new diagram named **mydiag2**. Drag and drop the **applicants** data set to the diagram workspace.
- From the Sample tab, add a Data Partition node to the diagram workspace and connect it to the **applicants** data set. Using the Properties panel, choose a training set of **50%**, a validation set of **30%**, and a test set of **20%**. Set the partitioning method to **stratified**. This ensures that the good/bad odds are the same in the training set, the validation set, and the test set.
- From the Model tab, add a Regression node to the diagram workspace and connect it to the Data Partition node. Accept the default settings and run the Regression node by right-clicking it and selecting **Run**. Inspect the output.
- From the Model tab, add a Neural Network node to the diagram workspace. Connect it to the Data Partition node and run it.
- From the Model tab, add a Decision Tree node to the diagram workspace. Connect it to the Data Partition node and run it.

- f. From the Assess tab, add a Model Comparison node to the diagram workspace and connect it to the Regression node, the Neural Network node, and the Decision Tree node. Run the Model Comparison node and inspect the results.
- 1) Which model gives the best ROC curve on the test set? Is this the same on the training and validation set?
 - 2) Which model gives the lowest misclassification rate on the test set? Is this the same on the training and validation set? (Hint: from the menu above, choose View, Model, Statistics Comparison)
- g. Right-click on the cumulative lift chart and select **Data Options....** Set the role of the **CAPC** variable to **Y** so as to obtain CAP/Lorenz curves for the training, validation and test sets.
- h. We are now ready to score new input data. We will assume that the data set **score** contains new applicants that need to be scored. Create a new data source for this data set which is also available in the mydata library. Make sure the variables have the right levels (see above) and set the Role of the data to **Score** in the properties panel. Drag and drop the **score** data to the diagram workspace.
- i. From the Assess tab, add a Score node to the diagram workspace. Connect the Model Comparison node to the Score node. Also connect the **score** data to the score node. Run the score node and inspect the results. You can do this by clicking the button next to the Exported data property in the properties panel. Browse the **score** data. The columns predicted:good_bad=good and predicted:good_bad=bad give the probability that the customer is good and bad, respectively.

Your workspace should now look as follows:



12. Estimating Decision Trees in SAS Enterprise Miner

- a. Create a new diagram named **mydiag3**. Drag and drop the **applicants** data set to the diagram workspace.
- b. From the Sample tab, add a Data Partition node to the diagram workspace and connect it to the **applicants** data set. Choose a training set of **70%** and a validation set of **30%** (no test set). Set the partitioning method to **stratified**. This ensures that the good/bad odds are the same in the training set and the validation set.
- c. From the Model tab, add a Decision Tree node to the diagram workspace and connect it to the Data Partition node. Inspect the Properties panel of the Decision Tree node. The SAS implementation of decision trees finds binary or multiway splits for nominal, ordinal, and interval inputs. The splitting criteria and other options to determine the method of tree construction can be chosen. The options include a mix of the features of C4.5\See5 (entropy splitting), CART (gini splitting), and CHAID (ProbChisq splitting). The Variance and ProbF criteria are used for interval target variables (such as **LGD**). Accept the default settings.
- d. If you want a binary yes/no tree, you can accept the default value of the option Maximum Branch, which is **2**.
- e. Set the maximum depth of the tree to **15**.

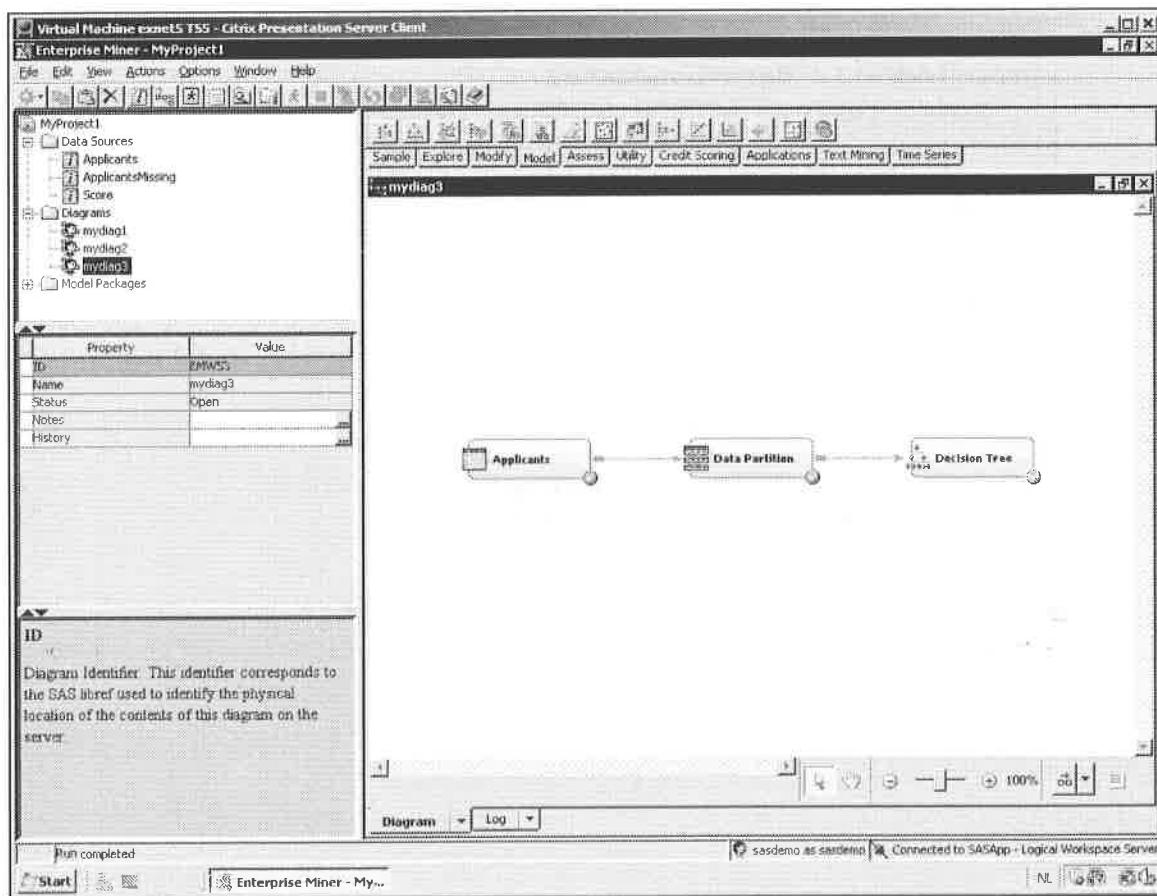
The Leaf Size option specifies the smallest number of training observations that a leave can have.

- f. The Number of Rules option controls the number of competitor splits reported for each node. The best competitor split has the second highest value for the entropy reduction compared to the optimal split. Accept the suggested setting for this option.
 - g. Regarding the Number of Surrogate Rules option, a *surrogate rule* is a rule that closely agrees with the action of the primary split. The degree of agreement is defined as the proportion of training observations that the two rules assign the same branch. When the main splitting rule relies on an input whose value is missing (for example, because it was not in the training set), the first surrogate rule is used. If this one is also missing, the second surrogate rule is used. Set the number of surrogate rules to **5**.
 - h. Sometimes, many splits are possible for a given variable. In order to limit the exhaustive search and possibly use a heuristic procedure, the Exhaustive option can be set (default = 5000). Accept the suggested setting for this option.
 - i. The Node Sample option sets an upper limit on the number of observations used to determine a split (by default 20000). Accept the suggested setting for this option.
- The Method option in the Subtree subpanel controls how the best subtree will be selected. When set to **assessment**, the smallest tree with the best assessment value on the validation set will be selected.
- j. The Assessment Measure (also in the Subtree subpanel) is used to select the best tree based on the validation set. Set this option to **Misclassification**.

k. Inspect some of the other options. Run the Decision Tree node and browse the results. Inspect the results carefully by looking at the following:

- The nodes of the tree are colored according to the impurity of the target variable. (Dark nodes indicate low entropy, whereas lighter nodes indicate high entropy.) Display the entire tree by right-clicking on the plot and then selecting **View \Rightarrow Fit to Page**.
- The tree map gives a compact graphical display of the tree in vertical orientation. The node width is proportional to the number of observations in the node and the color again corresponds to the impurity.
- The fit statistics provide a numerical summary of model performance.
- The leaf statistics display the percentage of goods in each leaf. The leaf with the highest percentage of goods on the training data has a leaf index equal to 1. Select the first leaf index in this plot, and see how the corresponding leaf in the tree window is highlighted.
- The evolution of the misclassification error on the validation and training set is viewed by selecting **View \Rightarrow Model \Rightarrow Subtree Assessment Plot**. Make sure that the Y axis shows the misclassification error and not the average squared error. You might have to change the graph properties for this.
- The English rule corresponding to the tree is one rule for each path from the root node to a leaf node. You can set the rules by selecting **View \Rightarrow Model \Rightarrow English rules**.
- The cumulative lift chart.

Your window now resembles the following:





The Decision Tree node can also be run in interactive mode. To do this, click the button next to the **Interactive** property in the Properties panel.

Multiclass Classification

13. Multiclass Cumulative Logistic Regression

- Start Base SAS and open the file **multiclasslogisticregressionstart.sas**, which is located in the d:\workshop\winsas\bb4C directory.
- Start by converting the **rating** variable to numerical values as follows:

```
data bondrate2;
  set bondrate;
  if rating='AAA' then rating=1;
  if rating='AA' then rating=2;
  if rating='A' then rating=3;
  if rating='BAA' then rating=4;
  if rating='BA' then rating=5;
  if rating='B' then rating=6;
  if rating='C' then rating=7;
run;
```

- Inspect the **bondrate2** data set in the **Work** library to see if the conversion went well.
- Estimate a cumulative logistic regression classifier using PROC LOGISTIC as follows:

```
proc logistic data=bondrate2;
  model rating= LOPMAR LFIXCHAR LGEARRAT LTDCAP LLEVER
    LCASHRAT LACIDRAT LCURRAT LRETURN LASSLTD;
  output out=cumlogout predprobs=individual;
run;
```

- e. Inspect both the output and the generated output data set **cumlogout**.

With the option PREDPROBS=INDIVIDUAL in the OUTPUT statement, you can ask for the individual estimated probabilities per rating for each observation (denoted as IP_x in the output data set **cumlogout**).

```

Virtual Machine exnet5 FSS - Citrix Presentation Server Client
SAS - [multiclasslogisticregressionstart]
File Edit View Tools Run Solutions Window Help
Results Results: Logistic: The SAS System
data bondrate2;
set bondrate;
if rating='AAA' then rating=1;
if rating='AA' then rating=2;
if rating = 'A' then rating=3;
if rating = 'BAA' then rating=4;
if rating= 'BA' then rating=5;
if rating= 'B' then rating=6;
if rating= 'C' then rating=7;
run;

proc logistic data=bondrate2;
model rating=LOPMAR LFIXCHAR LGEARRAT LTDCLP LLEVER
LCASHRAT LACIDRAT LCURRAT LRECTURN LASSLTD;
output out=cumlogout predprobs=individual;
run;

```

NOTE: 117 Lines Submitted.

Start Enterprise Miner - MyPro... SAS - [multiclasslogisticregressionstart] Results Viewer - SAS Out...

Measuring Scorecard Performance

14. Scorecard Performance

- a. Consider the following scorecard performance data set:

Score	Actual Good/Bad
100	Bad
110	Bad
120	Good
130	Bad
140	Bad
150	Good
160	Bad
170	Good
180	Good
190	Bad
200	Good
210	Good
220	Bad
230	Good
240	Good
250	Bad
260	Good
270	Good
280	Good
290	Bad
300	Good
310	Bad
320	Good
330	Good
340	Good

- b. Draw the Kolmogorov-Smirnov Curve.
- c. Calculate the Kolmogorov-Smirnov statistic.
- d. Draw the ROC curve.
- e. Check that the scorecard performs better than a random scorecard.
- f. Draw the CAP/Lorenz curve.
- g. Again, check that the scorecard performs better than a random scorecard.

You can do this in Base SAS as follows (thanks to Jana Honnerova, Czech Republic):

```

data perf;
    input score actual_gb $;
    cards;
100 b
110 b
120 g
130 b
140 b
150 g
160 b
170 g
180 g
190 b
200 g
210 g
220 b
230 g
240 g
250 b
260 g
270 g
280 g
290 b
300 g
310 b
320 g
330 g
340 g
;

proc sql noprint;
    select count(*) into :goods from perf where actual_gb = "g";
    select count(*) into :bads from perf where actual_gb = "b";
quit;

```

(Continued on the next page.)

```
%let all = %eval(&goods + &bads);
data discrimstat;
  set perf;
  retain g b 0;
  if actual_gb = "b" then b+1;
  else g + 1;
  sens = (&goods - g) / &goods;
  spec = (b / &bads);
  neg_spec = 1 - spec;
  neg_sens=1-sens;
  KS=spec-neg_sens;
  percentile=_N_/25;
run;

goptions reset = all;
symbol i = join ci = red v = dot cv = blue;
proc gplot data = discrimstat;
  plot sens * neg_spec;
  label sens = "Sensitivity" neg_spec = "1 - Specificity";
  title 'ROC plot';
run;
quit;

goptions reset = all;
symbol1 i = join ci = red v = dot cv = blue;
symbol2 i = join ci = green v = dot cv = blue;
proc gplot data = discrimstat;
  plot neg_sens *score spec*score/overlay;
  title 'Kolmogorov Smirnov plot';
run;
quit;

goptions reset = all;
symbol i = join ci = red v = dot cv = blue;
proc gplot data = discrimstat;
  plot spec*percentile;
  title 'CAP plot';
run;
quit;
```

Input Selection

15. Chi-Squared Filter and Cramer's V

- Consider the **applicants** data set.
- Apply the Chi-squared filter and Cramer's V using PROC FREQ to rank the following inputs according to their predictive power: Marital status, Job, Checking account, Purpose, and Savings account.

The SAS code should look as follows:

```
proc freq data=mydata.applicants;
  table marital*good_bad /chisq;
run;
```

- Complete the following table:

	Chi-square	p-value	Cramer's V
Marital status			
Job			
Checking account			
Purpose			
Savings account			

16. Stepwise Logistic Regression

Consider the **applicants** data set. Perform backward, forward, and stepwise logistic regression and determine which inputs are considered important for these three methods. Use SLENTRY=0.10 and SLSTAY=0.05 to start and examine the impact on the results when you vary these parameters. (See the course handouts for the SAS code.)

```
proc logistic data=mydata.applicants;
  class checking history purpose savings employed marital coapp
    resident property age other housing;
  model good_bad=checking duration history purpose amount savings
    employed installp marital coapp resident property other <
    /selection=stepwise slentry=0.10 slstay=0.05;
run;
```

SAS Credit Scoring Nodes

17. Developing PD Scorecards Using the Credit Scoring Nodes in SAS Enterprise Miner

- Create a new diagram named **CSnodes**.
- Drag and drop the **applicants** data set to the diagram workspace.
- From the Modify tab, add a Transform variables node to the diagram workspace and connect it to the **applicants** data set. Select the **Transform variables** node, go to the Properties panel, and click the button next to the **SAS Code** property.

- d. Type the following code:

```
if good_bad='bad' then good_bad =1;
else good_bad =0;
```

This ensures that, in what follows, the bad customers are considered the *events*. Hence, the scorecard assigns low scores to the bad customers and high scores to the good customers. If you do not do this transformation, the scorecard assigns low scores to the good customers and high scores to the bad customers, which is less intuitive.

- e. From the Credit Scoring tab, add an Interactive Grouping node to the diagram workspace and connect it to the Transform Variables node. Inspect some of the properties of the Interactive Grouping node.

The Use Frozen Groupings option can be used to prevent automatic grouping from being performed when the node is run and previous groupings already exist. The Binning method option specifies the method to do pre-binning of interval variables and it enables both quantile and bucket binning. The interval grouping method and ordinal grouping method option enables you to specify the method to do the grouping:

- **Optimal criterion:** finds the best groupings based on the Tree Based Criterion options. In the Tree Based Criterion options, the splitting criterion can be set to Entropy or ProbChisq (Chi-squared analysis).
- **Quantile:** creates groups with approximately an equal amount of observations in each.
- **Monotonic event rate:** requests groups to be monotonic in the event rate.
- **Constrained optimal:** creates groups based on pre-defined constraints, which can be set using both the Constrained Optimal Options property and the Advanced Constrained Optimal property.

The Maximum Number of Groups specifies the maximum number of groups to be generated. The Significant Digits option enables you to specify the numerical precision up to which interval variables are grouped. If this value is 0, then the lower and upper values of an interval group are integers.

The Adjust WOE option enables you to deal with groups in which all observations have the same target value (for example, all good or all bad). The probabilities then become $p_goodattribute = (\text{number of goodattribute} + \text{adjustment factor}) / \text{number of goodtotal}$ and $p_badattribute = (\text{number of badattribute} + \text{adjustment factor}) / \text{number of badtotal}$. The adjustment factor can also be specified in the Properties panel.

The Variable Selection Method property specifies the statistic used for variable selection: either the Gini statistic or the Information Value. You can also specify the cut-off for both of these statistics.

- f. Run the Interactive Grouping node and inspect the results. Inspect the different output windows. The Event Rate Plots window presents event rate plots for each variable. An event rate plot presents, for each group of the variable, the number of events in that group divided by the total number of events (300, in this case). The Output Variables window shows which variables were removed and which were retained based on the specified statistic (in this case, information value > 0.1). The statistics plot presents the information values of all variables in a bar chart.

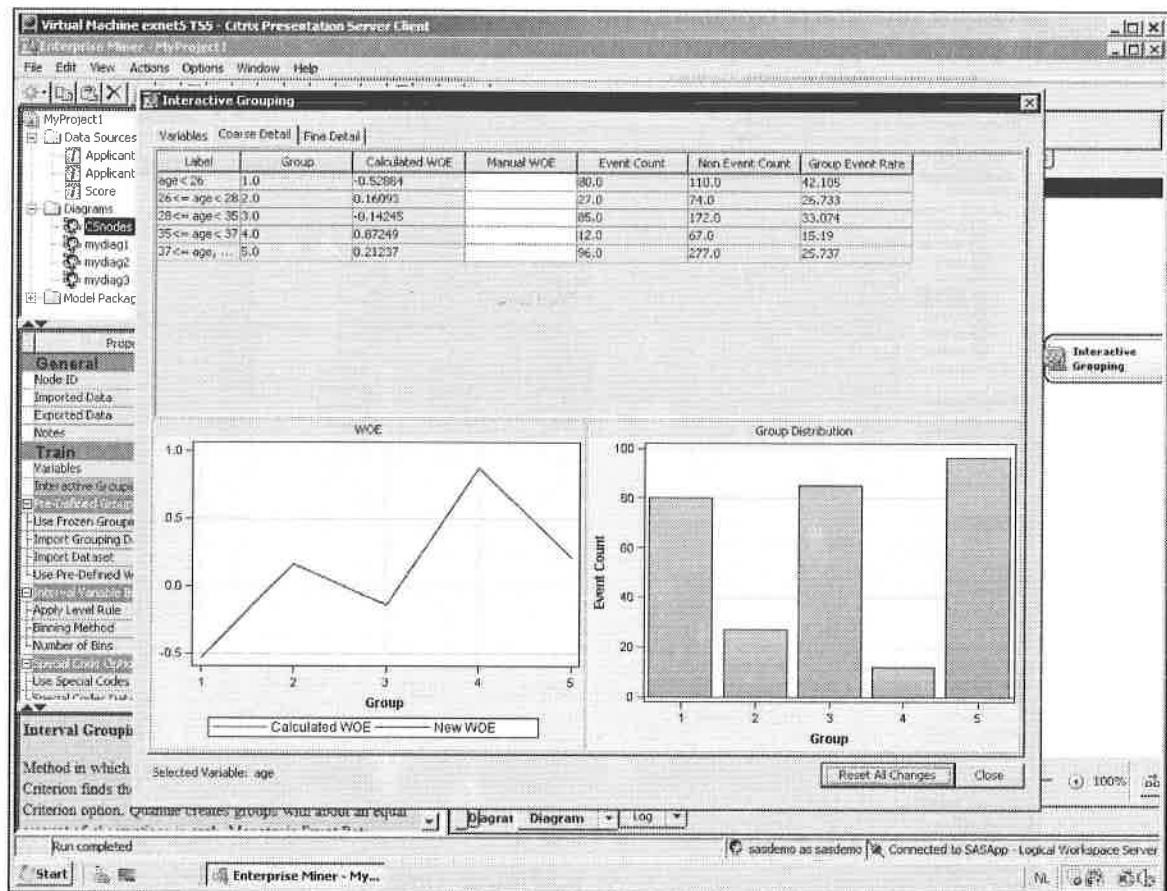
- g. The Interactive Grouping node can also be run in interactive mode. Close the Results window and select the **Interactive Grouping** node again. Click the button next to the **Interactive Grouping** property in the properties panel. The following window appears:

The screenshot shows the SAS Enterprise Miner interface with the title bar "Virtual Machine exnet5 TSS - Citrix Presentation Server Client" and the URL "pdcerx05065.exnet.sas.com". The main window displays the "Interactive Grouping" dialog box. The left pane shows the project structure under "MyProject" with nodes like "Data Sources", "Diagrams", and "CNodes". The right pane shows the "Variables" tab of the "Interactive Grouping" dialog. The table lists variables such as "checking", "history", "duration", "savings", "purpose", "amount", "property", "employed", "housing", "other", "marital", "instalip", "coapp", "coaddr", "job", "telephon", "resident", "age", "depends", and "foreign". Columns include Variable, Label, Pre-D..., Level, Exported Role, New Role, Original Gini, and Original Information... . The "Variables" tab is selected, and the "Diagram" tab is visible at the bottom.

Variable	Label	Pre-D...	Level	Exported Role	New Role	Original Gini	Original Information...
checking			ORDINAL	Input	Input	41.554	0.666
history			NOMINAL	Input	Input	25.311	0.292
duration			INTERVAL	Input	Input	23.451	0.251
savings			ORDINAL	Input	Input	19.617	0.191
purpose			NOMINAL	Input	Input	21.751	0.166
amount			INTERVAL	Input	Input	15.595	0.134
property			NOMINAL	Input	Input	17.066	0.113
employed			NOMINAL	Rejected	Rejected	16.164	0.066
housing			NOMINAL	Rejected	Rejected	13.436	0.033
other			NOMINAL	Rejected	Rejected	9.619	0.056
marital			NOMINAL	Rejected	Rejected	10.465	0.045
instalip			INTERVAL	Rejected	Rejected	8.677	0.026
coapp			NOMINAL	Rejected	Rejected	2.667	0.016
coaddr			INTERVAL	Rejected	Rejected	4.31	0.01
job			NOMINAL	Rejected	Rejected	4.137	0.0990
telephon			NOMINAL	Rejected	Rejected	3.905	0.0660
resident			NOMINAL	Rejected	Rejected	3.236	0.0040
age			INTERVAL	Rejected	Rejected	0.714	0.0010
depends			INTERVAL	Rejected	Rejected	0.238	0.0
foreign			NOMINAL	Rejected	Rejected	0.0	0.0

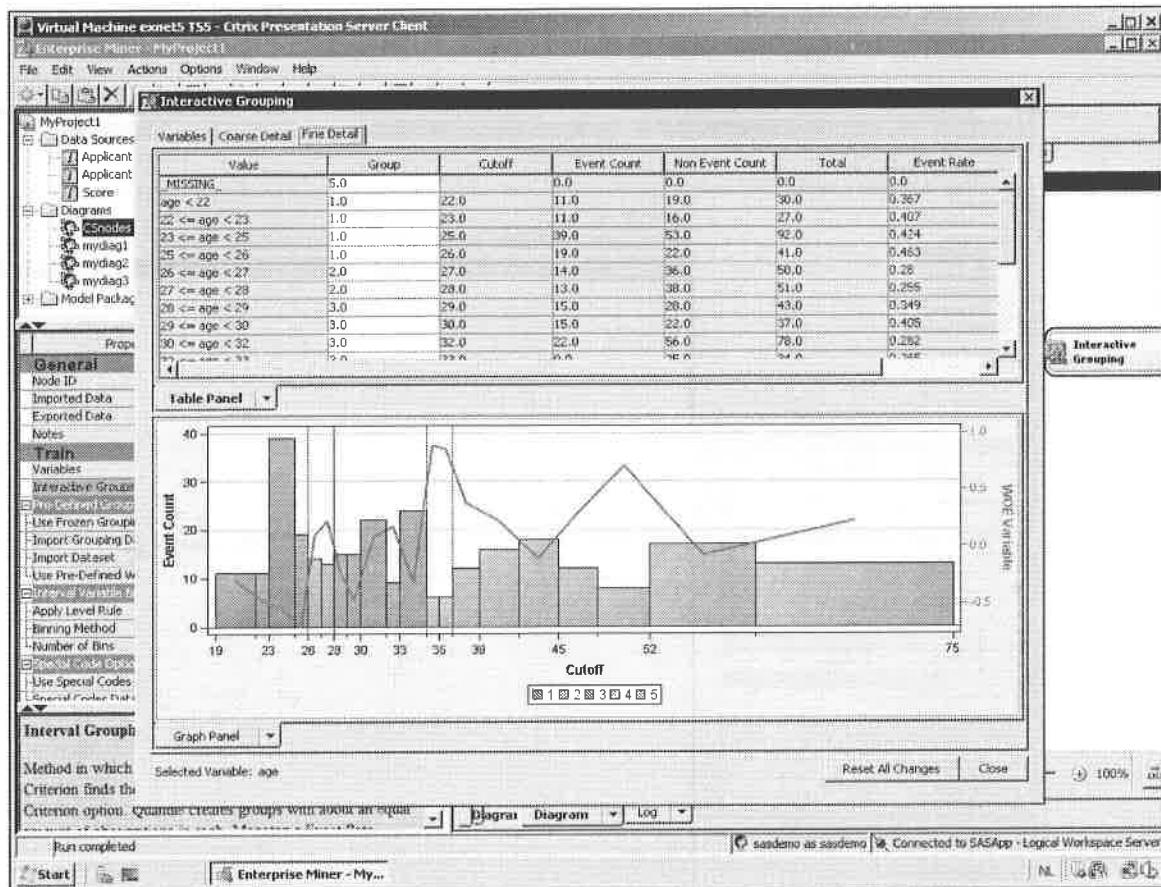
- j. The Variables tab can be used to navigate between the different variables. It also reports the original Gini and original Information Value that SAS obtained. Select the **age** variable and click the **Select** button below.

k. Click the Coarse Detail tab.



Here you see how the groups were constructed, a weights of evidence, and a group distribution plot.

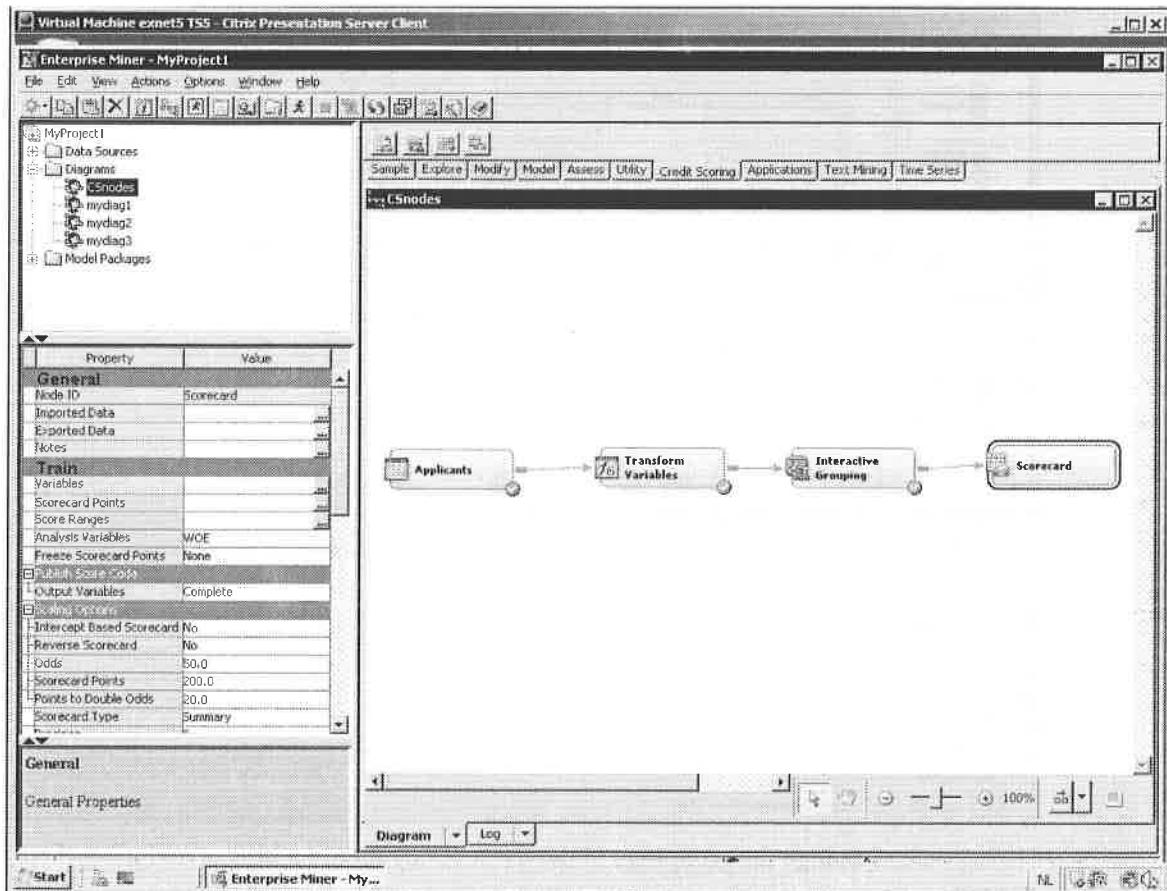
h. Go to the **Fine Detail** tab.



Here, you can change the groupings of the variables by merging bins or assigning them to other groups.

- 1) Select some values in the table above by holding down the SHIFT key and right-clicking to either merge or assign them to other groups. Make some changes in the grouping of the **age** variable.
- 2) Go back to the Variables tab and select the **Purpose** variable.
- 3) Explore the Coarse Detail and Fine Detail tabs.
- 4) Assign the X value of the purpose variable to another group.
- 5) Inspect the impact on the Information Value and compare it with the Information Value that SAS obtained (to be seen in the Original Information Value column of the Variables tab).

- i. From the Credit Scoring tab, add a Scorecard node to the diagram workspace and connect it to the Interactive Grouping node. This node estimates a logistic regression classifier using the weights of evidence variables. Your display should look as follows:



- j. Inspect the properties of the Scorecard node. The Odds option is the event/non-event odds that correspond to the score value that you specify in the Scorecard Points option, which is the sum of the score points across the various attributes of a customer at the given odds.
- 1) Set the Odds to **50** (default), the Scorecard Points to **600**, and the Points to Double Odds to **20** (default).
 - 2) Set the Scorecard Type option to **Detailed**. A summary scorecard displays the attribute names and the scorecard points for each attribute. A detailed scorecard contains additional information such as group number, weight of evidence, event rate, percentage of population, and regression coefficient for each attribute.
 - 3) Set the number of buckets to **20**. (This is the number of intervals in which the score range will be divided in a Gains table or Trade-off chart.)
 - 4) Set the revenue of an accepted good to **1000**, the cost of an accepted bad to **5000**, the Current approval rate to **70**, and the Current event rate to **2.5**.
 - 5) Set the Generate Characteristic Analysis option to **Yes**.

- k. In the Adverse Characteristics Option section, you specify one of the following scores that is used to identify adverse characteristics:

Weighted average score

The *weighted average score* of a characteristic (variable) is the weighted average of the scorecard points of observations in the groups. For example, suppose that there are three groups for the variable **age**, and the average score points of the groups are 36, 42, and 45, with 20, 30, and 40 observations in each of the groups, respectively. The weighted average score of **age** then becomes $(36*20+42*30+45*40)/90=42$.

Neutral score

The *neutral score* is the score points when weights of evidence is equal to 0 (that is, odds=1, where the probabilities of good and bad are equal for each attribute). Because the equation for calculating the score points is $\text{score} = \log(\text{odds}) * \text{factor} + \text{offset}$, the value of the score equals the offset. Hence, the neutral score equals the offset and is the same for each characteristic (variable).

- l. It is required by law to explain why an applicant is rejected. You can do this by comparing the actual value of a variable to the weighted average score or neutral score in order to identify characteristics that are deemed adverse. For example, the following table lists the weighted average score and actual value of variables of an applicant who was rejected, and the characteristics are listed based on the values of difference in ascending order. In this example, the characteristics that are deemed adverse in the order from most severe to least severe are **own_rent**, **income**, **age**, and **education**.

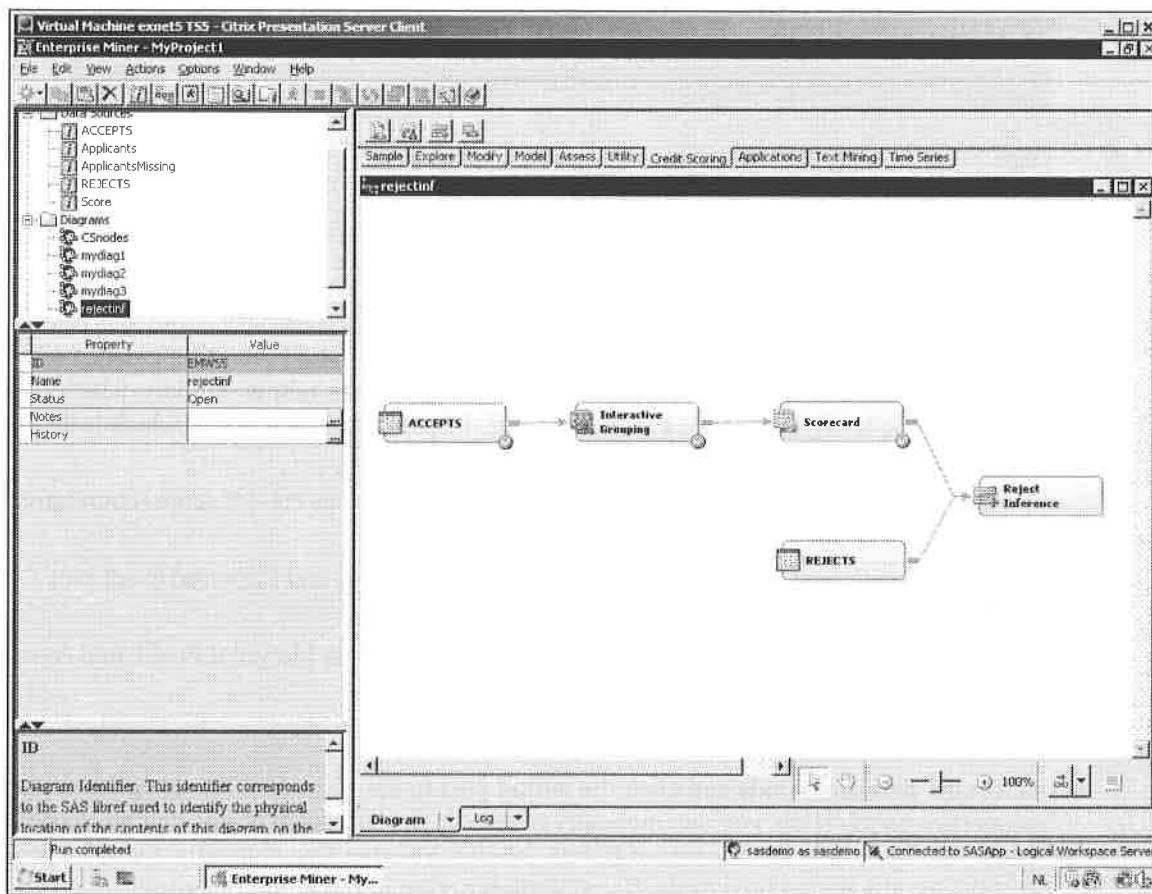
Characteristic	Weighted Average Score	Actual Value	Difference
own_rent	64	57	-7
income	54	52	-2
age	33	35	2
education	59	65	6

Set the option to **Weighted Average Score**.

- m. Run the Scorecard node and inspect the output. Inspect the scorecard in the Scorecard window. Inspect the KS-Statistic, AUC, and Gini index of the Scorecard in the Fit Statistics window. Notice that the Gini coefficient equals $2 * \text{AUC} - 1$. Inspect the output window, which contains the characteristic analysis for the various variables. Inspect the other output, which is available from the View menu above, such as the following:
- Event Frequency charts: depicting event frequencies versus cut-off score (cumulative or not, percentage wise or not)
 - Strength statistics: Kolmogorov-Smirnov plot, ROC plot, and Captured event plot (CAP/Lorenz curve)
 - Trade-Off plots: displays Cumulative Event Rate, Average Marginal Profit, and Average Total Profit
- n. Close the Results window.
- o. Select the Scorecard node and click the button next to the Scorecard Points option in the properties panel. Here, you can manually overwrite the scorecard points that the scorecard has generated by editing the Scorecard Points column. You can also manually adapt the score ranges by clicking the button next to the Score Ranges option in the properties panel.

18. The Reject Inference Node

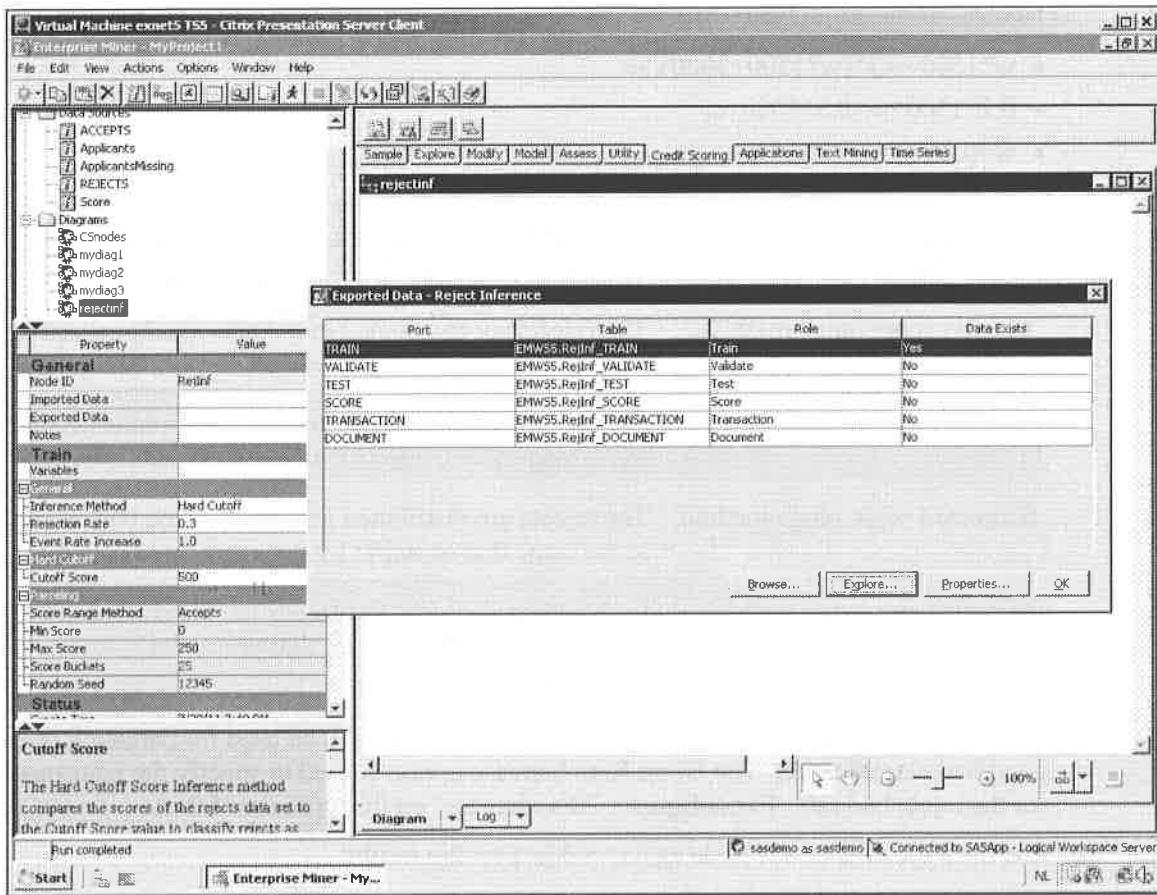
- Create a new diagram named **rejectinf**.
- Create two new data sources for the **accepts** and **rejects** data sets (both in the **mydata** library) respectively.
 - For the **accepts** data set, set the role of the **GB** variable to **target** and its level to **binary**. Set the role of the **_FREQ_** variable to **rejected**.
 - Accept the default settings for the other variables. For the **rejects** data set, accept the default settings for all variables.
- Drag and drop the **accepts** data set to the diagram workspace.
- From the Credit Scoring tab, add an Interactive Grouping node to the diagram workspace and connect it to the **accepts** data set. Run the Interactive Grouping node. Use the default settings.
- From the Credit Scoring tab, add a Scorecard node to the diagram workspace and connect it to the Interactive Grouping node. Set the Scorecard Points option to **600**, the Odds option to **50** (default), and the Points to Double Odds to **20** (default). Run the Scorecard node.
- Drag and drop the **rejects** data set to the diagram workspace. Set the role of this data set to **Score**. Use the Properties panel.
- From the Credit Scoring tab, add a Reject Inference node to the diagram workspace. Connect the Scorecard node and the **rejects** data set to the Reject Inference node. Your workspace should resemble the following:



- h. Inspect the properties of the Reject Inference node.
- 1) Set the Inference Method option to **Hard Cutoff**.
 - 2) Set the Cutoff Score option to **500**.
 - 3) Set the Rejection Rate option to **0.3** (default). The rejection rate determines the weighting of the rejects in the augmented data set (cf. infra). When the rejection rate is lower, the rejects are less important in the augmented data set. Here is the equation:

$$\text{Bad Rate}_{\text{augmented}} = w_1 * \text{Bad Rate}_{\text{accepts}} + w_2 * \text{Bad Rate}_{\text{rejects}}$$
For example, if reject rate = 0.3 then the following is true:

$$\text{Bad Rate}_{\text{augmented}} = 0.7 * \text{Bad Rate}_{\text{accepts}} + 0.3 * \text{Bad Rate}_{\text{rejects}}$$
- i. Run the Reject Inference node and inspect the results. The Output Data window shows the training data event rate, the inferred data event rate, and the augmented data event rate, taking into account the rejection rate. The Interval Variables window shows the distribution of the scorecard points and the event/non-event probabilities for the **accepts** and **inferred** data sets.
- j. Go back to the Properties panel and click the button next to the **Exported Data** property. Explore the new training data set that was created.



- k. In the Sample Properties panel, set the value of the Fetch Size option to **Max**.

- 1) Click the **Apply** button.
- 2) Inspect the data.

Various new columns were added to the data set. The Unnormalized Into: GB column looks at the columns Predicted: GB=1 and Predicted: GB=0, and assigns the observation to the class with the highest probability. The Into:GB column has the same values as the Unnormalized Into: GB column. The Good/Bad and From: GB columns have the same values. Both columns assign observations according to the scorecard points and the cutoff score chosen before. In this case, observations are assigned 0 (good customer) if the score is above 500, and 1 (bad customer) otherwise. Notice that there is also a column, **_FREQ1_**, added to the data set. This column contains the weight for the accepts and the inferred. In this case, the accepts get a weight of 1. The weight for the rejects w is then computed as follows: $w*1500 = \text{reject_rate} * (w*1500 + 3000)$.

You have the following:

- $\text{reject_rate} = 0.3$
- weighted accepts=3000 and unweighted rejects=1500

Hence,

- $w*1500 = 0.3 * (w*1500 + 3000)$, or
- $0.7 * 1500 * w = 0.3 * 3000$, or
- $w = 0.8571$

- l. Close the Results window, and run the Reject Inference node again using the parcelling method. Set the Inference Method option to **parcelling**. The Score Range Method option enables you to specify the range of scores to be bucketed as follows:

Accepts score range method	The rejects are distributed into equal-sized buckets based on the score range of the accepts data set.
Rejects score range method	The rejects are distributed into equal-sized buckets based on the score range of the rejects data set.
Scorecard score range method	The rejects are distributed into equal-sized buckets based on the score range that is output by the augmented data set.
Manual score range method	In the manual setting, the rejects are distributed into equal sized buckets based on the values that you specify for the Min Score and Max Score options.

- m. The Score Buckets option specifies the number of buckets to be used for parcelling during the good/bad classification. The Event Rate Increase option is used to specify the adjustment value for the reject bad rate. To configure a 20% increase, set this option to **1.2**.
- n. Run the Reject Inference node again, and inspect the results.

19. Weighted Average LGD

Do this exercise with a simple calculator (for example, the calculator available in Windows or Microsoft Excel).

Suppose you are given the following data:

Year 1: 30 defaults of \$50 with average loss of 10%

Year 2: 20 defaults of \$80 with average loss 70% and

40 defaults of \$100 with average loss of 60%

Complete the following table:

Long run LGD	Default count averaging	Exposure weighted averaging
Default weighted averaging	50	50
Time weighted averaging	50	50

Which weighted LGD would you choose according to Basel II?

20. Modeling LGD Using the Beta Distribution

- a. Load the data **LGD.txt** into SAS as follows:

```
proc import out= work.lgd
    datafile= "D:\workshop\winsas\bb4c\LGD.txt"
    dbms=tab replace;
getnames=yes;
datarow=2;
run;
```

- b. The first 13 columns of the data set are ratios that will be used to predict the LGD, which is the last column. Inspect the distribution of the **LGD** variable using SAS/INSIGHT. Select **Solutions** \Rightarrow **Analysis** \Rightarrow **Interactive Data Analysis** from the menu.

- Does the distribution look skewed?
- Does it look Gaussian?

- c. Split the 506 observations into a training set of 337 and a holdout set of 169 observations as follows:

```
data training holdout;
set LGD;
if 1 <= _N_ <= 337 then output training;
else output holdout;
run;
```

- d. Inspect both the **training** and the **holdout** data sets to see whether they were successfully created. Start by estimating an ordinary linear regression model without transforming the target variable as follows:

```
proc reg data=training outest=outests;
  LGDOLS: model LGD= ratio1-ratio13;
run;
```

- e. Look at the *p*-values of the regression to see the most important inputs. Compute the predictions on the holdout set as follows:

```
proc score data=holdout score=outests out=preds type=parms;
  var ratio1-ratio13;
run;
```

- f. Inspect the **preds** data set. Visualize the performance of the linear regression model using a scatter plot in SAS/INSIGHT. Select **Solutions** \Rightarrow **Analysis** \Rightarrow **Interactive Data Analysis** from the menu and create a scatter plot. Compute the correlation between the actual and predicted LGD as follows:

```
proc corr data=preds;
  var LGD LGDOLS;
run;
```

Report the correlation in the table at the end of this exercise.

- g. Compute the mean squared error (MSE) on the holdout set as follows:

```
data MSEtemp;
  set preds;
  MSEterm=(LGD-LGDOLS)**2;
run;

proc means data=MSEtemp;
  var MSEterm;
run;
```

Report the MSE in the table at the end of this exercise.

- h. You are ready to do the second regression where you assume that the **LGD** variable has a Beta distribution and transform it. Compute the mean and the variance of the **LGD** variable as follows:

```
proc means data=training;
  var LGD;
  output out=meanstats mean=mu var=sigmasq;
run;
```

- i. Estimate the alpha and beta parameters of the Beta distribution as follows:

```
data betaparams;
  set meanstats;
  alpha=(mu*mu*(1-mu)/sigmasq)-mu;
  beta=alpha*(1/mu -1);
run;
```

- j. Inspect the **betaparams** data set.

- k. Transform the **LGD** variable to a normal distributed variable as follows:

```
data transformedtraining;
  if _N_=1 then set betaparams;
  set training;
  newLGD=probit(cdf('BETA',LGD,alpha,beta));
run;
```

- l. Inspect the distribution of the **newLGD** variable using SAS/INSIGHT. Select **Solutions** \Rightarrow **Analysis** \Rightarrow **Interactive Data Analysis** from the menu and inspect the distribution of the **newLGD** variable.

Does it look more symmetric and Gaussian now?

- m. Estimate a linear regression model on the transformed variable as follows:

```
proc reg data=transformedtraining outest=outests2;
  LGDBETA: model newLGD= ratio1-ratio13;
run;
```

- n. Look at the *p*-values of the regression to see the most important inputs. Are these the same as in the previous model? Compute the predictions on the **holdout** data set as follows:

```
proc score data=holdout score=outests2 out=transpreds type=parms;
  var ratio1-ratio13;
run;
```

- o. Transform the predictions as follows:

```
data pred2;
  if _N_=1 then set betaparams;
  set transpreds;
  LGDpred=betainv(cdf('NORMAL',LGDBETA),alpha,beta);
run;
```

- p. Visualize the performance of the transformed linear regression model using a scatter plot in SAS/INSIGHT. Select **Solutions** \Rightarrow **Analysis** \Rightarrow **Interactive Data Analysis** from the menu and make a scatter plot of **LGD** versus **LGDpred**. Compute the correlation between the actual and predicted **LGD** as follows:

```
proc corr data=pred2;
  var LGD LGDpred;
run;
```

Report the correlation in the table at the end of this exercise.

- q. Compute the MSE of the transformed linear regression model as follows:

```
data MSEtemp2;
  set pred2;
  MSEterm=(LGD-LGDpred)**2;
run;

proc means data=MSEtemp2;
  var MSEterm;
run;
```

Report the MSE in the table below.

	Correlation	MSE
Linear regression		
Beta linear regression		

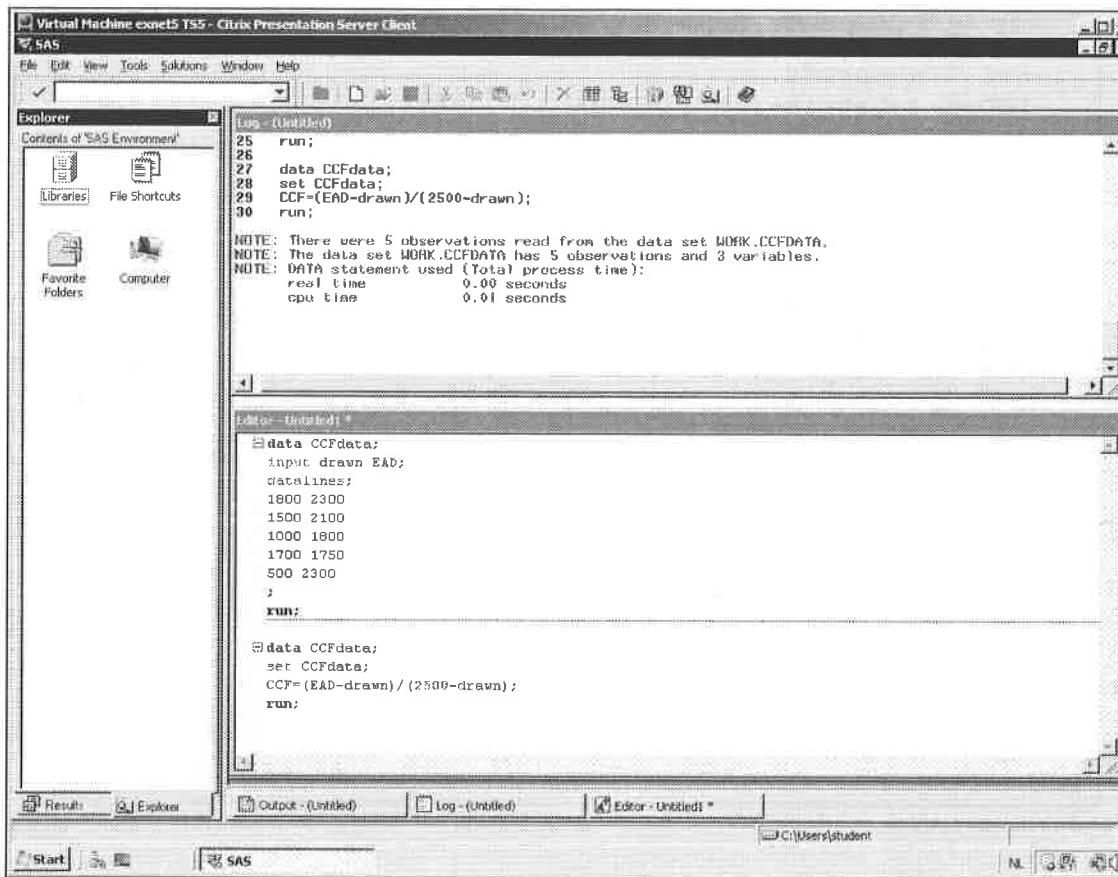
According to your findings, which model gives the best performance?

21. EAD Modeling

Consider a credit card portfolio with the following characteristics for five customers that defaulted in the past:

EAD 1 month prior to default	EAD at time of default
1800	2300
1500	2100
1000	1800
1700	1750
500	2300

Each credit card has a credit limit of \$2,500. Compute the credit conversion factor (CCF) for each customer in the portfolio.



The screenshot shows a SAS software interface with two main windows. The top window is titled 'Log - (Untitled)' and contains the following SAS code and its output:

```

Virtual Machine exnet5 TS5 - Citrix Presentation Server Client
File Edit View Tools Solutions Window Help
Log - (Untitled)
25 run;
26
27 data CCFdata;
28 set CCFdata;
29 CCF=(EAD-drawn)/(2500-drawn);
30 run;

NOTE: There were 5 observations read from the data set WORK.CCFDATA.
NOTE: The data set WORK.CCFDATA has 5 observations and 3 variables.
NOTE: DATA statement used (Total process time):
      real time          0.00 seconds
      cpu time           0.01 seconds

```

The bottom window is titled 'Editor - (Untitled)*' and contains the same SAS code, showing the input data and the calculation of CCF:

```

Editor - (Untitled)*
data CCFdata;
input drawn EAD;
datalines;
1800 2300
1500 2100
1000 1800
1700 1750
500 2300
;
run;

data CCFdata;
set CCFdata;
CCF=(EAD-drawn)/(2500-drawn);
run;

```

22. Backtesting at Level 0

- a. Compute a system stability index (SSI) for the following data:

Score range	Actual%	Training%
0-169	7%	8%
170-179	8%	10%
180-189	7%	9%
190-199	9%	13%
200-209	11%	11%
210-219	11%	10%
220-229	10%	9%
230-239	12%	10%
240-249	11%	11%
250+	14%	9%

- b. Enter the data in SAS as follows:

```
data populationstab;
input range$ actual training;
datalines;
0-169 0.07 0.08
170-179 0.08 0.10
180-189 0.07 0.09
190-199 0.09 0.13
200-209 0.11 0.11
210-219 0.11 0.10
220-229 0.10 0.09
230-239 0.12 0.10
240-249 0.11 0.11
250+ 0.14 0.09
;
run;
```

- c. Compute the intermediate values as follows:

```
data SSIdata;
set populationstab;
temp1=actual-training;
temp2=log(actual/training);
index=temp1*temp2;
run;
```

- d. Now compute the SSI as follows:

```
proc means data=SSIdata sum;
run;
```

- Do you observe a difference?
- What traffic light color would you assign?

- e. Play with the numbers in the **populationstab** data and observe the impact on the SSI.

The screenshot shows the SAS software interface. The Editor window contains the following SAS code:

```

78
79 proc means data=SSIdata sum;
80 run;

Editor - Untitled1 *
data populationstab;
  input range$ actual training;
  datalines;
  0-169 0.07 0.08
  170-179 0.08 0.10
  180-189 0.07 0.09
  190-199 0.09 0.13
  200-209 0.11 0.11
  210-219 0.11 0.10
  220-229 0.10 0.09
  230-239 0.12 0.10
  240-249 0.11 0.11
  250+ 0.14 0.09
;
run;

data SSIdata;
  set populationstab;
  temp1=actual-training;
  temp2=log(actual/training);
  index=temp1*temp2;
  run;

proc means data=SSIdata sum;
run;

```

The Log window shows the execution results:

```

real time      0.01 seconds
cpu time       0.00 seconds

```

23. Backtesting at Level 2: Binomial Test

- a. Consider the following data from an IRB rating system:

Rating Category	Estimated PD	Number of Observations	Number of Observed Defaults
A	2%	1000	17
B	3%	500	20
C	7%	400	35
D	20%	200	50

- b. Compute the critical values according to the normal approximation of the binomial test and see whether there are any significant differences between realized default rates and estimated PDs.

- c. Enter the data in Base SAS as follows:

```
data ratings;
  input rating$ estimatedPD nrobs nrdefaults;
  lines;
A 0.02 1000 17
B 0.03 500 20
C 0.07 400 35
D 0.20 200 50
;
run;
```

- d. Compute the actual default rates and the critical values as follows:

```
data binomialtest;
  set ratings;
  actualdefaultrate=nrdefaults/nrobs;
  criticalvalue=probit(0.99)*
    ((estimatedPD*(1-estimatedPD)/nrobs)**0.5)+estimatedPD;
run;
```

- e. Inspect the binomial test data set. Check every rating and see whether the actual default rate exceeds the critical value or not. Alter the **nrdefaults** variable and investigate the impact on the results.

24. Backtesting at Level 2: Hosmer-Lemeshow Test

- a. Consider the following data from an IRB rating system (same as in previous exercise):

Rating Category	Estimated PD	Number of Observations	Number of Observed Defaults
A	2%	1000	17
B	3%	500	20
C	7%	400	35
D	20%	200	50

- b. Conduct a Hosmer-Lemeshow test to see whether the observed data agree with the predicted PDs.
 c. Create a SAS data set in Base SAS as follows:

```
data ratings;
  input rating$ estimatedPD nrobs nrdefaults;
  lines;
A 0.02 1000 17
B 0.03 500 20
C 0.07 400 35
D 0.20 200 50
;
run;
```

- d. Create a new data set containing all summation terms of the Hosmer-Lemeshow statistic as follows:

```
data chisquaredvaluetemp;
  set ratings;
  chisquaredterm= (nrobs*estimatedPD-
    nrdefaults) **2 / (nrobs*estimatedPD*(1-estimatedPD)) ;
run;
```

- e. Inspect the **chisquaredvaluetemp** data set. Now sum all these terms to obtain the value of the test statistic. Use PROC MEANS.

```
proc means data=chisquaredvaluetemp;
  var chisquaredterm;
  output out=chisquaredvalue sum=sum_chisquaredterm;
run;
```

- f. Inspect the **chisquaredvalue** data set. Compute the *p*-value of the test statistic as follows:

```
data HLstatistic;
  set chisquaredvalue;
  pvalue=1-CDF('chisquared',sum_chisquaredterm,_FREQ_);
run;
```

- g. The **_FREQ_** variable represents the degrees of freedom. Inspect the **HLstatistic** data set.
- What does the *p*-value tell you?
 - Does the observed data differ significantly from the estimated PDs?
 - How does your conclusion relate to the outcome of the binomial test?
 - Change the **ratings** data to alter the answer to the previous question.

25. Backtesting at Level 2: Traffic Light Indicator Approach

Suppose you are given the following output from a Traffic Light Indicator approach to backtesting:

PD	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av
	0,15%	0,09%	0,37%	0,74%	0,72%	1,91%	3,32%	5,84%	7,40%	17,07%	2,98%
DR	Baa1	Baa2	Baa3	Ba1	Ba2	Ba3	B1	B2	B3	Caa-C	Av
1993	0,00%	0,00%	0,00%	0,83%	0,00%	0,76%	3,24%	5,04%	11,29%	28,57%	3,24%
1994	0,00%	0,00%	0,00%	0,00%	0,00%	0,59%	1,88%	3,75%	7,95%	5,13%	1,88%
1995	0,00%	0,00%	0,00%	0,00%	0,00%	1,76%	4,35%	6,42%	4,06%	11,57%	2,51%
1996	0,00%	0,00%	0,00%	0,00%	0,00%	0,00%	1,17%	0,00%	3,28%	13,99%	0,78%
1997	0,00%	0,00%	0,00%	0,00%	0,00%	0,47%	0,00%	1,54%	7,22%	14,67%	1,41%
1998	0,00%	0,31%	0,00%	0,00%	0,62%	1,12%	2,11%	7,55%	5,52%	15,09%	2,83%
1999	0,00%	0,00%	0,34%	0,47%	0,00%	2,00%	3,28%	6,91%	9,63%	20,44%	3,35%
2000	0,28%	0,00%	0,97%	0,94%	0,63%	1,04%	3,24%	4,10%	10,68%	19,65%	3,01%
2001	0,27%	0,27%	0,00%	0,51%	1,38%	2,93%	3,19%	11,07%	10,38%	34,45%	5,48%
2002	1,26%	0,72%	1,78%	1,58%	1,41%	1,58%	2,00%	6,81%	6,86%	29,45%	3,70%
Av	0,26%	0,17%	0,42%	0,53%	0,54%	1,36%	2,46%	5,76%	8,76%	20,9%	3,05%

The backtesting was done according to a one-tailed binomial test with respect to the reference PD given in the second row of the table. The colors are coded as follows:

Green	No difference at 10% level
Yellow	Difference at 10% level but not at 5%
Orange	Difference at 5% level but not at 1%
Red	Difference at 1% level

- How do you interpret the results from this backtesting exercise?
- What action would you suggest?

26. Benchmarking

- a. Consider the following data set containing ratings of 10 obligors provided by three institutions:

	Institution 1	Institution 2	Institution 3
Obligor 1	B	B	A
Obligor 2	A	A	C
Obligor 3	A	A	C
Obligor 4	D	B	B
Obligor 5	D	D	D
Obligor 6	A	A	C
Obligor 7	C	C	A
Obligor 8	B	A	B
Obligor 9	A	A	A
Obligor 10	D	D	D

Which two institutions provide the most similar ranking of obligors? Use Spearman's rho and Kendall's tau to determine this.

- b. Type the data in SAS as follows:

```
data ratings;
    input inst1$ inst2$ inst3$;
datalines;
B B A
A A C
A A C
D B B
D D D
A A C
C C A
B A B
A A A
D D D
;
run;
```

- c. Transform the data to numerical values as follows:

```

data ratings2;
  set ratings;
  if inst1='A' then do
    inst1num=1;
  end;
  if inst1='B' then do
    inst1num=2;
  end;
  if inst1='C' then do
    inst1num=3;
  end;
  if inst1='D' then do
    inst1num=4;
  end;
  if inst2='A' then do
    inst2num=1;
  end;
  if inst2='B' then do
    inst2num=2;
  end;
  if inst2='C' then do
    inst2num=3;
  end;
  if inst2='D' then do
    inst2num=4;
  end;
  if inst3='A' then do
    inst3num=1;
  end;
  if inst3='B' then do
    inst3num=2;
  end;
  if inst3='C' then do
    inst3num=3;
  end;
  if inst3='D' then do
    inst3num=4;
  end;
run;

```

- d. Compute Spearman's rho and Kendall's tau as follows:

```

proc corr data=ratings2 spearman kendall;
  var inst1num inst3num;
run;

```

- e. Complete the following table:

	Spearman's rho	Kendall's tau-b
Institution 1 versus Institution 2		
Institution 2 versus Institution 3		
Institution 1 versus Institution 3		

Which two institutions give the most similar ratings?

- f. Draw a histogram as follows:

```
data histo;
  set ratings2;
  hist1vs2=inst1num-inst2num;
  hist1vs3=inst1num-inst3num;
  hist2vs3=inst2num-inst3num;
run;

proc univariate data=histo nopolish;
  histogram hist1vs3 / cfill=blue midpoints=-4 to 4 by 1;
run;
```

Using a histogram, which two institutions give the most similar ratings?

- g. Change the numbers in the **ratings** data set at the start of the exercise and look at the impact on Spearman's rho and Kendall's tau and on the histograms.

27. Low Default Portfolios

Assume that you have a portfolio with zero defaults and three ratings: A (100 obligors), B (150 obligors), and C (80 obligors). You can assume that defaults are independent. Calculate the most prudent estimate of PD_A , PD_B , PD_C , assuming the following significance levels:

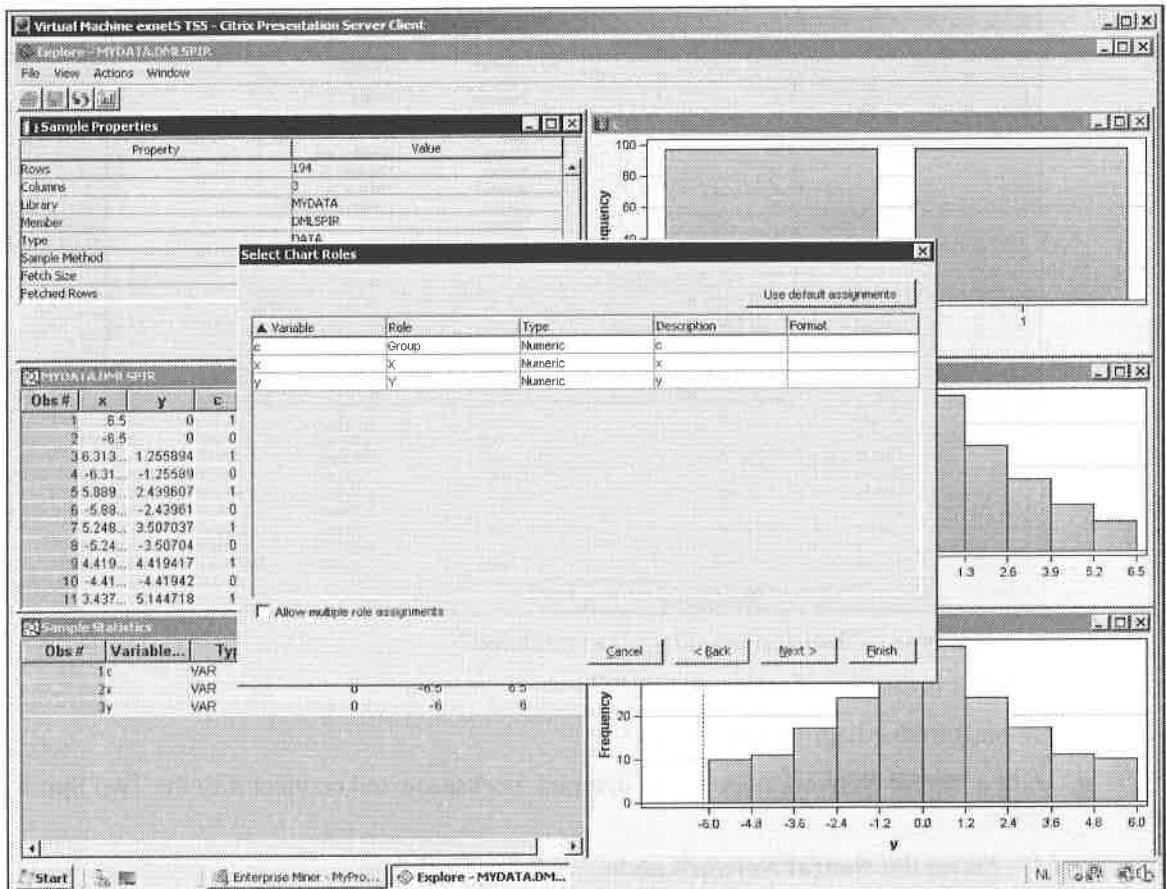
	$\alpha=0.90$	$\alpha=0.99$
PD_A		
PD_B		
PD_C		

```
%let sig=0.99;
%let nA=100;
%let nB=150;
%let nC=80;

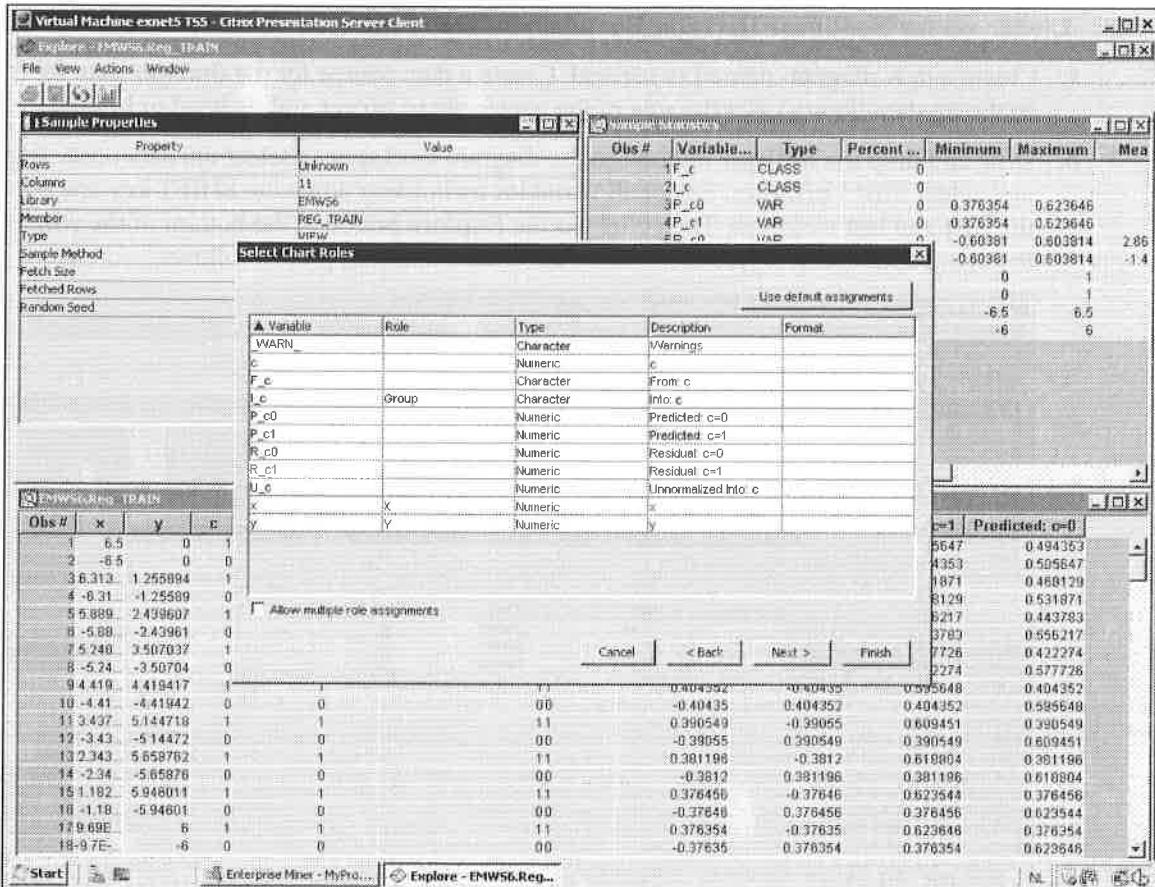
data LDP;
PDA=1- (1-&sig) **(1/(&nA+&nB+&nC));
PDB=1- (1-&sig) **(1/(&nB+&nC));
PDC=1- (1-&sig) **(1/(&nC));
run;
```

28. Linear versus Nonlinear Decision Boundary

- Create a new diagram named **twospiral**. Create a data source for the **dmlspir** data set (available in the **mydata** library). Set the role of the **c** variable to **target** and its level to **binary**.
- Drag and drop the **dmlspir** data set to the diagram workspace. Select the data node. Right-click and select **Edit Variables**. Select all variables by holding down the SHIFT key and clicking on the first and last variables. Then click on the **Explore** button at the bottom of the window. From the menu above, select **Actions** \Rightarrow **Plot** and create a scatter plot as follows:

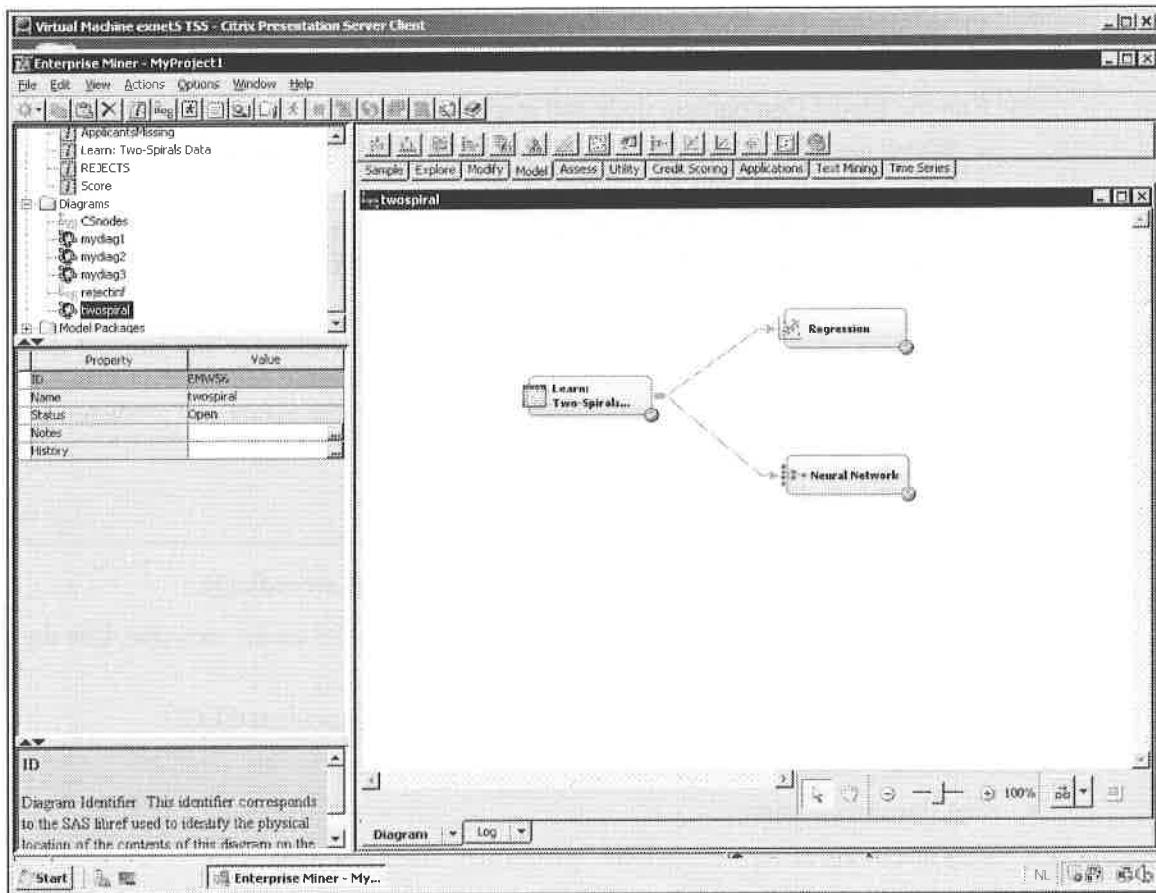


- Inspect the generated scatter plot. You clearly see the two spirals.
- Go back to the diagram workspace.
- Add a Regression node to the diagram workspace and connect it to the Two Spiral Data node.
 - Run the Regression node by right-clicking it and selecting **Run**.
 - Inspect the results of the Regression node.
 - Select the **Regression** node.
 - From the Properties panel, click the button next to the **Exported Data** property.
 - Select the **train** data and click the **Explore** button.
 - From the menu, select **Actions** \Rightarrow **Plot** and create a scatter plot. Set the role of the **X** variable to **X**, the role of the **Y** variable to **Y**, and the role of the **I_C** variable to **Group**.
 - Inspect the scatter plot.



- What type of decision boundary was modeled?
 - What is the classification accuracy?
- f. Go back to the diagram workspace.
- g. Add a Neural Network node to the diagram workspace and connect it to the Two Spiral Data node.
- 1) Select the **Neural Network** node.
 - a) In the Properties panel, click the button next to the **Network** property.
 - b) Set Number of Hidden Units to **40**.
 - c) In the Properties panel, click the button next to the Optimization property and set Maximum Iterations to **500**.
 - d) Run the Neural Network node by right-clicking it and selecting **Run**.
 - 2) Inspect the results of the Neural Network node. Select the **Neural Network** node.
 - a) From the Properties panel, click the button next to the **Exported Data** property.
 - b) Select the **train** data and click the **Explore** button. From the menu, select **Actions** \Rightarrow **Plot** and create a scatter plot. Set the role of the **X** variable to **X**, the role of the **Y** variable to **Y**, and the role of the **I_C** variable to **group**. Inspect the scatter plot.
 - What type of decision boundary has been modeled?
 - What is the classification accuracy?

- h. Vary the number of hidden neurons and inspect the impact on the decision boundary. Also estimate a decision tree with a Gini splitting criterion and inspect the decision boundary. Your window should resemble the following:



29. Neural Networks for Modeling LGD

- Create a new diagram named **NeuralLGD**.
 - Create a data source for the **LGDSAS** data set (available in the **mydata** library).
 - Set the role of the **LGD** variable to **target**.
 - Keep the level of all variables to **Interval** as suggested.
 - Explore the distribution of the **LGD** variable.
- Drag and drop the **LGDSAS** data set to the diagram workspace.
- From the Sample tab, add a Data Partition node to the diagram workspace and connect it to the **LGDSAS** data node. Accept the suggested training set, validation set, and test set percentages.
- Add four neural network nodes to the diagram workspace having respectively 2, 3, 5, and 10 hidden units. (You can set this by clicking the button next to the **Network** property in the Properties panel.)
 - For each Neural Network node, in the Properties panel set the Model Selection Criterion to **Average Error**.
 - Connect the Neural Network nodes to the Data Partition node.

- e. From the Assess tab, add a Model Comparison node to the diagram workspace and connect it to the four Neural Network nodes.
 - 1) Set the Selection Statistic property to **Mean Squared Error** and the Selection Table property to **Validation**. This node selects the neural network with the lowest average squared error on the validation set.
 - 2) Run the Model Comparison node and inspect the results.
 - Which neural network was selected by the Model Comparison node? Inspect the various types of errors on both the validation and the test set.
 - What is the mean squared error on the test set? Compare this with what you got when analyzing the **LGD** data with linear and beta-transformed linear regression in a previous exercise.
- f. From the Utility tab, add a SAS Code node to the diagram workspace and connect it to the neural network with the lowest validation set error. Enter the following code in the Code Editor property of the SAS Code node to compute the correlation between the actual and predicted LGDs:

```
proc corr data=EMWS.Neural_Test;
  var LGD P_LGD;
run;
```



You might have to change the name of the data set accordingly.

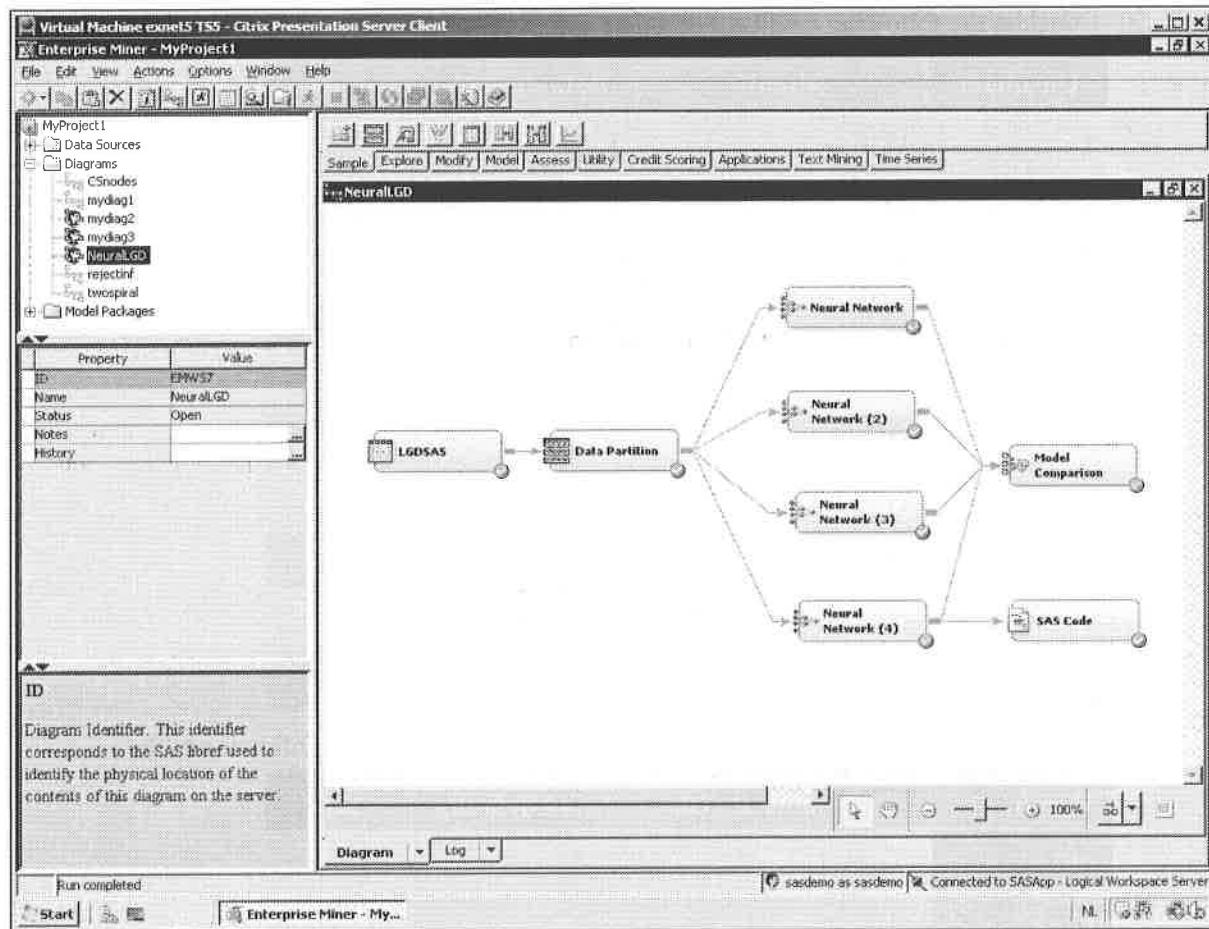
- g. Contrast the reported correlation with the ones reported in the earlier exercise. Use the linear regression approaches.
- h. Create a scatter plot that depicts the actual LGD versus the predicted LGD.

At this point, you tried three approaches to modeling LGD:

- using linear regression
- using linear regression with a Beta transformed target variable
- using neural networks.

Overall, which one gives the best performance?

Your display should resemble the following:



30. Kaplan Meier Analysis 1

- a. Consider the following credit scoring data set:

Customer	Month	Default (0=censored obs.)
1	17	1
2	27	1
3	30	0
4	32	1
5	43	1
6	43	1
7	47	1
8	52	1
9	52	0
10	52	0

- b. Compute the Kaplan Meier estimates manually by completing the following table:

Month	Customers at risk	Nr. of defaults	S(t)
<17			
17			
27			
30			
32			
43			
47			
52			

- c. Also compute the Kaplan-Meier estimates in BASE SAS and contrast them with your results. Plot the Kaplan-Meier curve.

The screenshot shows the SAS software interface. The title bar reads "Virtual Machine.exe 5 TSS - Citrix Presentation Server Client" and "SAS". The menu bar includes File, Edit, View, Tools, Run, Solutions, Window, and Help. The left sidebar has a "Results" section with a tree view showing "Lifetest: The SAS System". The main area has two windows: "Log - (Untitled)" and "Editor - Untitled1".

Log - (Untitled)

```
70 run;
71
72 proc lifetest data=survival1;
73 time month*censor(0);
74 run;

NOTE: The LOGLOG transform is used to compute the confidence limits for the quartiles of the survival distribution. To suppress using this transform, specify CONFTYPE=LINEAR in the PROC LIFETEST statement.
NOTE: PROCEDURE LIFETEST used (Total process time):
      real time           1.45 seconds
      cpu time            0.31 seconds
```

Editor - Untitled1

```
data survival1;
input month censor;
datalines;
17 1
27 1
30 0
32 1
43 1
43 1
47 1
52 1
52 0
52 0
;
run;

proc lifetest data=survival1;
time month*censor(0);
run;
```

At the bottom, there are tabs for Results, Output, Log, Editor, and Results Viewer. The status bar shows "C:\Users\student" and "In 11, Col 5".

31. Kaplan Meier Analysis 2

- a. Consider the tab-delimited data set **input.txt**. This is a credit scoring survival analysis data set with the following attributes:

Nr	Name	Explanation
1	Censor	1: customer defaulted; 0: good customer
2	Censore	1: customer paid loan back early; 0: customer did not pay back loan early
3	Open	For good customers: time of last observation For bad customers: time of default For customers that paid back early: time of early pay back
4	Age	Age of applicant
5	Amount	Amount of Loan
6	Currahd	Years at current address
7	Curremp	Years with current employer
8	Custgend	Gender of customer
9	Depchild	Number of dependent children
10	Freqpaid	Frequency of salary payments (for example, weekly or monthly)
11	Homephon	Has home phone or not
12	Insprem	Amount of insurance premium
13	Loantype	Type of loan (single or joint)
14	Marstat	Marital status
15	Term	Term of loan
16	Homeowns	Has home or not
17	Purpose	Purpose of loan

The data set can thus be used for predicting both customer default and early repayment.

- b. Type the data in SAS. Use the following SAS code:

```
data creditsurv;
  infile 'D:\workshop\winsas\bb4c\input.txt' delimiter=' ';
  input censor censore open age amount currahd curremp custgend
        depchild freqpaid homephon insprem loantype marstat term
        homeowns purpose;
run;
```

- c. Estimate a Kaplan-Meier survival curve for predicting default using the following SAS code:

```
proc lifetest data=creditsurv plots=(s) graphics;
  time open*censor(0);
  symbol1 v=none;
run;
```

The PLOTS=(s) option also provides a graphical representation of the KM curve. The output contains a line for each observation. Censored observations are starred. The second column gives the KM estimates. No KM estimates are reported for the censored times. The column labeled **Failure** is only 1 minus the KM estimate, which is the probability of default prior to the specified time. The fourth column is an estimate of the standard error of the KM estimate, obtained by Greenwood's formula (Collet 1994). This estimate can be used to construct confidence intervals. The column labeled **Number Failed** is only the cumulative number of customers that defaulted before or at the time point considered. The last column, labeled **Number left**, is the number of customers at risk during the considered time point. Below the main table, you find the estimated 75th, 50th, and 25th percentiles (labeled Quartiles). For example, the 25th percentile is the smallest event time, such that the probability of defaulting earlier is bigger than 0.25. Because in this case the survival probabilities are never lower than 0.9, no values are reported for these percentiles.

- d. Look at the KM graph. Why is the graph flat during the first three months? If you want to have the KM estimates at specific time points, you can use the TIMELIST= option as follows:

```
proc lifetest data=creditsurv plots=(s) graphics
  timelist=1,2,3,4,5,6,7,8,9,10;
  time open*censor(0);
  symbol1 v=none;
run;
```

- e. Suppose you want to test whether customers owning a home or not have the same survival curve. The null hypothesis then becomes $H_0: S_1(t)=S_2(t)$ for all t . PROC LIFETEST provides three test statistics for doing this:

- the log-rank test (also known as the Mantel-Haenzel test)
- the Wilcoxon test
- the likelihood-ratio statistic, which is based on the assumption of exponentially distributed event times

You can ask for these tests as follows:

```
proc lifetest data=creditsurv plots=(s) graphics;
  time open*censor(0);
  symbol1 v=none;
  strata homeowners;
run;
```

Is the difference significant? Also look at the graph with the two survival curves.

- f. Conduct the same test statistics to check whether **gender** has an impact on the survival curve. Also check whether the fact that a person is younger or older than 30 has an impact on the survival curve.

Hint: Use **strata age (30)**; in the code.

- g. Repeat the Kaplan Meier analysis for predicting early repayment using the censore indicator.

32. Parametric Survival Analysis

- a. Create a plot of $-\log(S(t))$ versus t for predicting default. Also create a plot of $\log[-\log(S(t))]$ versus t for predicting default. Do the plots support the assumption of an exponential or Weibull distribution of the survival times?

```
proc lifetest data=creditsurv plots=(s,ls,lls) graphics;
  time open*censor(0);
  symbol1 v=none;
run;
```

- b. Use PROC LIFEREG to estimate a parametric survival analysis model (for predicting default), assuming exponentially distributed survival times. The independent variables are **age**, **amount**, **curradd**, **curremp**, **custgend**, **depchild**, **freqpaid**, **homephon**, **insprem**, **loantype**, **marstat**, **term**, and **homeowns**. Be sure to make dummies for the **custgend**, **freqpaid**, **homephon**, **loantype**, **marstat**, and **homeowns** variables using the CLASS statement. Inspect the generated output. For a quantitative variable (for example, **age** and **amount**), you can use the formula $100(e^\beta - 1)$ to estimate the percent increase in the expected survival time for each one-unit increase in the variable, holding the other variables constant. Calculate this number for the variables **age** and **depchild**.

```
proc lifereg data=creditsurv;
  class custgend freqpaid homephon loantype marstat homeowns;
  model open*censor(0)=age amount curradd curremp custgend
    depchild freqpaid homephon insprem loantype marstat term
    homeowns / dist=exponential;
run;
```

- Is the time at current address important for predicting the survival time?
 - Is the gender important for predicting the survival time?
- c. Use PROC LIFEREG to estimate a parametric survival analysis model assuming Weibull, log-normal, and generalized gamma distributed survival times. Provide the log-likelihood of each of these models in the following table:

Distribution	Log-likelihood
Exponential	
Weibull	
Log-normal	
Generalized gamma	

Which model gives the best likelihood?

- d. Compare the various distributions using a chi-squared likelihood ratio test statistic as follows:
 $-2 \ln(L_{red}/L_{full}) = -2 \ln(L_{red}) + 2\ln(L_{full})$.

```
data LLdata;
  chi2testvalue=-2*(-5325)+2*(-5198);
  pvalue=1-CDF('chisquared',chi2testvalue,1);
run;
```

Models	Chi-Squared Likelihood Ratio Statistic	Degrees of Freedom	p-value
Exponential versus Weibull			
Exponential versus generalized gamma			
Weibull versus generalized gamma			
Log-normal versus generalized gamma			

- e. Using the OUTPUT statement, predicted survival times can be generated for all observations in the data set. If you want to have a point estimate for the survival time of an individual observation, it makes sense to use the median survival time, that is, t for which $S(t)=0.5$. This can be coded as follows:

```
proc lifereg data=creditsurv;
  class custgend freqpaid homephon loantype marstat homeowners;
  model open*censor(0)=age amount curradd curremp custgend
    depchild freqpaid homephon insprem loantype marstat term
    homeowners / dist=exponential;
  output out=preds p=median std=s;
run;

proc print data=preds;
  var open censor _prob_ median s;
run;
```

When inspecting the data set **preds**, it becomes obvious that most of the predicted survival times are much bigger than the **open** variable. This is due to the following:

- heavy censoring of the data
 - limited predictive power of the attributes
 - the choice of the median survival time $S(t)$ (Compare with setting a cut-off in classification models.)
- f. Repeat the parametric survival analysis for predicting early repayment using the **censore** indicator. Are the findings the same?

33. Proportional Hazards Regression

- a. Estimate a proportional hazards model for the **creditsurv** data set for predicting default using the interval variables **age**, **amount**, **curradd**, and **curremp**.

```
proc phreg data=creditsurv;
  model open*censor(0)=age amount curradd curremp;
run;
```

- b. The first part of the output gives information about testing the null hypothesis: $\text{Beta}=0$. The null hypothesis is that all β coefficients are 0. Three chi-squared test statistics are given. Is the model significant?

- c. The individual parameter estimates are then depicted. The Wald statistic can be computed by squaring the ratio of each coefficient to its estimated standard error, $(\beta/s(\beta))^2$, and has a chi-squared distribution with 1 degree of freedom. The risk ratio can be computed as $\exp(\beta)$. With the RISKLIMITS option, you can request confidence limits for the estimated hazard ratios.
- If age increases with 1, what is the percentage increase or decrease of the hazard?
 - If age increases with 1, what is the impact on the survival probability?
- d. By default, SAS uses the Breslow approximation to estimate the parameters. Contrast the estimated parameters with those obtained by using the Efron approximation. (Use TIES=EFRON.) Is there a big difference?
- e. You can also use PROC PHREG to generate predictions for individual observations. Start with creating a test set as follows:

```
data testset;
  input age amount curradd curremp;
  datalines;
36 3000 1 2
;
run;
```

- f. Estimate a proportional hazards model as follows:

```
proc phreg data=creditsurv;
  model open*censor(0)=age amount curradd curremp / ties=efron;
  baseline out=preds covariates=testset survival=s lower=lcl
           upper=ucl / nomean;
run;

proc print data=preds;
run;
```

The BASELINE statement enables you to ask for survival predictions of individual observations. The NOMEAN option suppresses the output of survival estimates evaluated at the mean values of the covariates, which are otherwise included by default. The LOWER= and UPPER= options give 95-percent confidence intervals around the survival probability.

- g. PROC PHREG also enables you to do input selection using the SELECTION= option with all three methods: backward, forward, and stepwise. Perform all three types of input selection and look at the impact on the results. (Use SLENTRY=0.01 and SLSTAY=0.01.)
- h. Repeat the analysis above for predicting early repayment.

Appendix B Data Dictionary for the Applicants Data Set

B.1 Data Dictionary B-3

B.1 Data Dictionary

Variable	Role	Level	Meaning
Age	Input	Interval	Age in years
Amount	Input	Interval	Amount of loan
Checking	Input	Ordinal	Status of existing checking account: 1: < 0 DM; 2: 0 to <200 DM; 3: >=200 DM/ salary assignments for at least 1 year; 4: no checking account
Coapp	Input	Nominal	Other debtors/guarantors: 1: none; 2: co-applicant; 3: guarantor
Depends	Input	Interval	Number of dependents
Duration	Input	Interval	Duration in months
Employed	Input	Nominal	Present employed since: 1: unemployed; 2: < 1 year; 3: 1 to < 4 years; 4: 4 to < 7 years; 5: >= 7 years
Existcr	Input	Interval	Number of existing credits at this bank
Foreign	Input	Binary	Foreign worker: 1: yes; 2: no
Good_bad	Target	Binary	Good/bad payer
History	Input	Nominal	0: no credits taken/all credits paid back duly; 1: all credits at this bank paid back duly; 2: existing credits paid back duly till now; 3: delay in paying off in the past; 4: critical account/other credits existing (not at this bank)
Housing	Input	Nominal	Housing: 1: rent; 2: own; 3: for free
Installp	Input	Interval	Installment rate in percentage of disposable income
Job	Input	Nominal	Job: 1: unemployed/unskilled – non-resident; 2: unskilled – resident; 3: skilled employee/official; 4: management/self-employed/highly qualified employee/officer

B-4 Appendix B Data Dictionary for the Applicants Data Set

Marital	Input	Nominal	Marital status: 1: male: divorced/separated; 2: female: divorced/separated/married; 3: male: single; 4: male: married/widowed; 5: female: single
Other	Input	Nominal	Other installment plans: 1: bank; 2: stores; 3: none
Property	Input	Nominal	Property: 1: real estate; 2: if not 1: building society savings agreement/life insurance; 3: if not 1/2: car or other, not in attribute 6; 4: unknown/no property
Purpose	Input	Nominal	Purpose of loan: 0: car (new); 1: car (used); 2: furniture/equipment; 3: radio/television; 4: domestic appliances; 5: repairs; 6: education; 7: vacation; 8: retraining; 9: business; X: others
Resident	Input	Interval	Date beginning permanent residence
Savings	Input	Ordinal	Savings account/bonds: 1: < 100 DM; 2: 100 to < 500 DM; 3: 500 to < 1000 DM; 4: >= 1000 DM; 5: unknown/no savings account
Telephon	Input	Binary	Telephone: 1: none; 2: yes, registered under the customer's name

Appendix C References

C.1 References C-3

C.1 References

- Allison, Paul D. 1999. *Logistic Regression Using the SAS® System: Theory and Application*. Cary, NC: SAS Institute Inc.
- Allison, Paul D. 2010. *Survival Analysis Using the SAS® System: A Practical Guide (second edition)*. Cary, NC: SAS Institute Inc.
- Altman, Edward, Andrea Resti, and Andrea Sironi, eds. 2005. *Recovery Risk: The Next Challenge in Credit Risk Management*. London: Risk Books.
- Araten, Michel, Michael Jacobs, Jr., and Peeyush Varshney. 2004. "Measuring LGD on Commercial Loans: An 18-Year Internal Study. *The RMA Journal*.
- Baesens B., et al. 2003. "Neural Network Survival Analysis for Personal Loan Data." *Proceedings of the Eighth Conference on Credit Scoring and Credit Control (CSCCVII'2003)*, Edinburgh, Scotland.
- Baesens, Bart, et al. 2003. "Benchmarking State of the Art Classification Algorithms for Credit Scoring." *Journal of the Operational Research Society* 54:627–635.
- Baesens, Bart, et al. 2003. "Using Neural Network Rule Extraction and Decision Tables for Credit-Risk Evaluation." *Management Science* 49:312–329.
- Baesens, Bart, et al. 2005. "Neural Network Survival Analysis for Personal Loan Data." *Journal of the Operational Research Society* (Special Issue on Credit Scoring) 59:1089–1098.
- Baesens, Bart. 2003. "Developing Intelligent Systems for Credit Scoring Using Machine Learning Techniques (Ph.D. dissertation, K.U.Leuven, Belgium).
- Banasik, John, Jonathan N. Crook, and Lyn C. Thomas. 1999. "Not if but when borrowers will default." *Journal of the Operational Research Society*.
- Banasik, John, Jonathan N. Crook, and Lyn C. Thomas. 2001. "Sample selection bias in credit scoring models." *Proceedings of the Seventh Conference on Credit Scoring and Credit Control (CSCCVII'2001)*, Edinburgh, Scotland.
- Bartlett, P.L. 1997. "For valid generalization, the size of the weights is more important than the size of the network." In *Advances in Neural Information Processing Systems 9*, eds. Mozer, M.C., Jordan, M.I., and Petsche, T., 134–140. Cambridge: MIT Press.
- Basel Committee on Banking Supervision. 2005. "Stress testing at major financial institutions: survey results and practice."
- Basel Committee on Banking Supervision. 2005. "Validation of low-default portfolios in the Basel II Framework." Newsletter No.6.
- Basel Committee on Banking Supervision. 2006. "International Convergence of Capital Measurement and Capital Standards: A Revised Framework Comprehensive."
- Basel Committee on Banking Supervision. 2006. "Studies on the Validation of Internal Rating Systems." BIS Working paper 14.
- Basel Committee on Banking Supervision. 2009. "Principles for sound stress testing practices and supervision."

- Benjamin, N., A. Cathcart, and K. Ryan. 2006. "Low default portfolios: A proposal for conservative estimation of default probabilities." Financial Services Authority.
- Bishop, C.M. 1999. *Neural Networks for Pattern Recognition*. New York: Oxford University Press.
- Breiman, Leo., et al. 1984. *Classification and Regression Trees*. Belmont, CA: Wadsworth International.
- Brown I., Mues C. 2011. "Regression Model Development for Credit Card Exposure at Default." Submitted.
- Cantor, A. 1997. *SAS Survival Analysis Techniques for Medical Research*. Cary, NC: SAS Institute Inc.
- Castermans G., Martens D., Van Gestel T., Hamers B., Baesens B. 2010. "An overview and framework for PDbacktesting and benchmarking." *Journal of the Operational Research Society* 61: 359–373.
- Cespedes, Juan Carlos Garcia. 2002. "Credit Risk Modeling and Basel II." *Algo Research Quarterly*. Volume 5, Number 1.
- Collet, D. 1994. *Modelling Survival Data in Medical Research*. London: Chapman & Hall.
- Committee of European Banking Supervisors (CEBS). 2005. "Guidelines on the implementation, validation and assessment of Advanced Measurement (AMA) and Internal Ratings Based (IRB) approaches." CP10 consultation paper.
- Cox, D.R. 1972. "Regression Models and Life Tables." *Journal of the Royal Statistical Society, Series B*.
- Cox, D.R. and D. Oakes. 1984. *Analysis of Survival Data*. London: Chapman and Hall.
- Cristianini N. and Taylor J.S. 2000. *An Introduction to Support Vector Machines and other kernel-based learning methods*. Cambridge University Press.
- Crowder, M.J. 2001. *Classical Competing Risks*. London: Chapman and Hall.
- Dwyer, D.W., A. Kocagil, and R. Stein. 2004. "The Moody's KMV EDF™ RiskCalc™ v3.1 Model Next-Generation Technology for Predicting Private Firm Credit Risk." White paper.
- Engelmann, B., E. Hayden, and D. Tasche. 2003. "Measuring the Discriminative Power of Rating Systems." Deutsche Bundesbank discussion paper.
- Fawcett, T. 2003. "ROC Graphs: Notes and Practical Considerations for Researchers." *HP Labs Tech Report HPL-2003-4*.
- Financial Services Authority (FSA). 2005. "CP05/3: Strengthening Capital Standards 2." FSA consultation paper.
- Financial Services Authority (FSA). 2006. "CP06/3: Strengthening Capital Standards 2." FSA consultation paper.
- Financial Services Authority (FSA). 2009. "Stress and Scenario Testing." FSA PS 09/20.
- Freed, N. and F. Glover. 1986. "Resolving Certain Difficulties and Improving the Classification Power of LP Discriminant Analysis Formulations." *Decision Sciences* 17:589–595.
- Fritz, S. 2003. "Validating Inputs and Outputs of the Regulatory and Economic Capital Allocation Process." Advanced IRB Forum presentation.

- Grablowsky, B.J. and W.K. Talley. 1981. "Probit and discriminant factors for classifying credit applicants: A comparison." *Journal of Economics and Business* 33:254–261.
- Gupton, G. and R. Stein. 2005. "LossCalc v2: Dynamic Prediction of LGD." Research paper.
- Hall, M. and L. Smith L. 1996. "Practical feature subset selection for Machine Learning." *Proceedings of the Australian Computer Science Conference* (University of Western Australia).
- Hand, D. and R.J. Till. 2001. "A simple generalization of the area under the ROC curve to multiple class classification problems." *Machine Learning* 45(2):171–186.
- Hand, D.J. 2003. "Crime, statistics, and behaviour." *Statistics, Science, and Public Policy VIII: Science, Ethics, and the Law*. Queen's University, Canada.
- Hand, D.J. and S. Jacka, eds. 1998. *Statistics in Finance*. London: Edward Arnold.
- Hand, D.J. and W.E. Henley. 1997. "Statistical classification methods in consumer credit scoring: A review." *Journal of the Royal Statistical Society Series A* 160 3:523–541.
- Hand, D.J., H. Mannila H., and P. Smyth. 2001. *Principles of Data Mining*. Cambridge: MIT Press.
- Hanley, J. A. and B. J. McNeil. 1982. "The meaning and use of area under the ROC curve." *Radiology* 143:29–36.
- Hartigan, J.A. 1975. *Clustering Algorithms*. New York: Wiley.
- Henley, W.E. and D.J. Hand. 1997. "Construction of a k-nearest neighbour credit scoring system." *IMA Journal of Mathematics Applied in Business and Industry* 8:305–321.
- Hong Kong Monetary Authority (HKMA). 2006. "Validating Risk Rating Systems under the IRB Approaches."
- Joseph, M.P. 2005. "Validation Framework for Basel II Internal Ratings-Based Systems." Commonwealth Bank of Australia.
- Kalbfleisch, D. and R.L. Prentice. 2003. *The Statistical Analysis of Failure Time Data*. New York: Wiley.
- Kelly, M.G. 1998. "Tackling Change and Uncertainty in Credit Scoring." (Unpublished Ph.D. thesis, The Open University, Milton Keynes, UK.)
- Kleinbaum, D. 1996. *Survival Analysis: A Self-Learning Text*. New York: Springer.
- Leow M, Mues C. 2011. Credit Risk Models for Mortgage Loans. Submitted.
- Loterman G, Brown I., Martens D., Mues C., Baesens B. 2011. "Benchmarking regression algorithms for loss given default modeling." *International Journal of Forecasting (forthcoming)*.
- Mangasarian, O.L. 1965. "Linear and Non-linear Separation of Patterns by Linear Programming." *Operations Research*. 13:444–452.
- Mars, M. 2003. "Internal ratings validation survey." *The RMA Journal*.
- Merton, R. C. 1974. "On the Pricing of Corporate Debt: The Risk Structure of Interest Rates." *Journal of Finance* 29:449–470.

- Moges H., Lemahieu W., Baesens B. 2011. "A Multidimensional Analysis of Data Quality for Credit Risk Management: New Insights and Challenges." Submitted.
- Monetary Authority of Singapore (MAS). 2002. "Credit Stress-Testing."
- Moody, J. and J. Utans. 1994. "Architecture Selection Strategies for Neural Networks: Application to Corporate Bond Rating Prediction." In *Neural Networks in the Capital Markets*, ed. Refenes, A.N., New York: Wiley.
- Moral G. 2006. "EAD Estimates for Facilities with Explicit Limits." In: Engelmann B, Rauhmeier R (Eds), *The Basel II Risk Parameters: Estimation, Validation and Stress Testing*, Springer, Berlin, 197–242.
- Ong, M. K. 2004. *The Basel Handbook: A Guide for Financial Practitioners*. London: Risk Books.
- Pluto, K. and D. Tasche. 2005. "Estimating Probabilities of Default for Low Default Portfolios." Working paper.
- Potts, W. 2004. *Neural Network Modelling*. SAS.
- Quinlan, J.R. 1993. *C4.5: Programs for Machine Learning*. San Francisco: Morgan Kauffman.
- Ripley, B.D. 1996. *Pattern Recognition and Neural Networks*. Cambridge University Press.
- Saunders, A. and L. Allen. 2002. *Credit Risk Measurement: New Approaches to Value at Risk and Other Paradigms*. New York: Wiley.
- Schölkopf, B. and A. Smola. 2002. *Learning with Kernels*. Cambridge: MIT Press.
- Schuermann, T. and Y. Jaffry. 2004. "Measurement, Estimation, and Comparison of Credit Migration Matrices." *Journal of Banking and Finance*.
- Stoyanov S. 2009. Application LGD Model Development. *Credit Scoring and Credit Control XI Conference*.
- Thomas, L. C., D. Edelman, and J. Crook. 2002. "Credit Scoring and Credit Control." Society for Industrial and Applied Mathematics (SIAM).
- Thomas, L.C. 2000. "A survey of credit and behavioural scoring; Forecasting financial risk of lending to consumers." *International Journal of Forecasting* 16:149–172.
- Thomas, L.C. 2009. *Consumer Credit Models: Pricing, Profit and Portfolios*. Oxford University Press.
- Thomas, L.C. and M. Stepanova. 2001. "PHAB scores: Proportional hazards analysis behavioural scores." *Journal of the Operational Research Society* 52:1007–16.
- Thomas, L.C. and M. Stepanova. 2002. "Survival analysis methods for personal loan data." *Operations Research* 50:277–289.
- Thomas, L.C., J. Ho, and W.T. Scherer. 2001. "Time will tell: Behavioural scoring and the dynamics of consumer risk assessment." *IMA Journal of Management Mathematics* 12:89–103.
- Treacy, W.F. and M. Carey. 1998. "Credit Risk Rating at Large U.S. Banks." *Federal Reserve Bulletin* 84:897–921.

- Van Gestel T., Martens D., Vanden Branden K., Baesens B. 2011. "A Practical Framework for Credit Risk Stress Testing." Submitted.
- Van Gestel, T. 2002. "From Linear to Kernel Based Methods in Classification, Modelling and Prediction." ESAT-SCD-SISTA. Katholieke Universiteit Leuven, Belgium.
- Van Gestel, T. and Baesens B. 2009. *Credit Risk Management: Basic Concepts: Financial Risk Components, Rating Analysis, Models, Economic and Regulatory Capital*. Oxford University Press.
- Van Gestel, T., et al. 2005. "Linear and nonlinear credit scoring by combining logistic regression and support vector machines." *Journal of Credit Risk* Volume 1, Number 4.
- Van Gestel, T., et al. 2006. "A process model to develop an internal rating system: sovereign credit ratings." *Decision Support Systems* 42:1131–1151.
- Vapnik, V. 1995. *The Nature of Statistical Learning Theory*. New York: Springer-Verlag.
- Vasicek, O. 1987. "Probability of Loss on Loan Portfolio." San Francisco: KMV Corporation.
- Vasicek, O. 1991. "Limiting Loan Loss Probability Distribution." San Francisco: KMV Corporation.
- Zurada, J.M. 1992. *Introduction to Artificial Neural Systems*. Boston: PWS Publishing.



Recommended SAS® Titles

Credit Risk Modeling Using SAS®

ISBN	Title	Price (U.S. Dollars)
SAS® Press		
978-0-470-16776-2	<i>Fair Lending Compliance: Intelligence and Implications for Credit Risk Management</i>	\$75.00
978-0-470-46168-6	<i>Credit Risk Assessment: The New Lending System for Borrowers, Lenders, and Investors</i>	\$49.95
978-0-471-75451-0	<i>Credit Risk Scorecards: Developing and Implementing Intelligent Score Credit Scoring</i>	\$49.95

Notes

- Prices are subject to change without notice.
- To order, please visit support.sas.com/bookstore.
- SAS documentation is available to search, browse, or print free online at: support.sas.com/documentation.

