



2013-2

Credit Scoring Using Machine Learning

Kenneth Kennedy

Dublin Institute of Technology

Follow this and additional works at: <http://arrow.dit.ie/sciendoc>

Recommended Citation

Kennedy, K. (2013). *Credit scoring using machine learning*. Doctoral thesis. Dublin Institute of Technology. doi:10.21427/D7NC7J.

This Theses, Ph.D is brought to you for free and open access by the Science at ARROW@DIT. It has been accepted for inclusion in Doctoral by an authorized administrator of ARROW@DIT. For more information, please contact yvonne.desmond@dit.ie, arrow.admin@dit.ie, brian.widdis@dit.ie.



This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License



Credit Scoring Using Machine Learning

by

Kenneth Kennedy

Supervisors: Dr. Brian Mac Namee

Dr. Sarah Jane Delany

Prof. Pádraig Cunningham



School of Computing

Dublin Institute of Technology

A thesis submitted for the degree of

Doctor of Philosophy

February, 2013

For Daniel.

Abstract

For financial institutions and the economy at large, the role of credit scoring in lending decisions cannot be overemphasised. An accurate and well-performing credit scorecard allows lenders to control their risk exposure through the selective allocation of credit based on the statistical analysis of historical customer data. This thesis identifies and investigates a number of specific challenges that occur during the development of credit scorecards. Four main contributions are made in this thesis.

First, we examine the performance of a number supervised classification techniques on a collection of imbalanced credit scoring datasets. Class imbalance occurs when there are significantly fewer examples in one or more classes in a dataset compared to the remaining classes. We demonstrate that oversampling the minority class leads to no overall improvement to the best performing classifiers. We find that, in contrast, adjusting the threshold on classifier output yields, in many cases, an improvement in classification performance.

Our second contribution investigates a particularly severe form of class imbalance, which, in credit scoring, is referred to as the low-default portfolio problem. To address this issue, we compare the performance of a number of semi-supervised classification algorithms with that of logistic

regression. Based on the detailed comparison of classifier performance, we conclude that both approaches merit consideration when dealing with low-default portfolios.

Third, we quantify the differences in classifier performance arising from various implementations of a real-world behavioural scoring dataset. Due to commercial sensitivities surrounding the use of behavioural scoring data, very few empirical studies which directly address this topic are published. This thesis describes the quantitative comparison of a range of dataset parameters impacting classification performance, including: (i) varying durations of historical customer behaviour for model training; (ii) different lengths of time from which a borrower's class label is defined; and (iii) using alternative approaches to define a customer's default status in behavioural scoring.

Finally, this thesis demonstrates how artificial data may be used to overcome the difficulties associated with obtaining and using real-world data. The limitations of artificial data, in terms of its usefulness in evaluating classification performance, are also highlighted. In this work, we are interested in generating artificial data, for credit scoring, in the absence of any available real-world data.

Declaration

I certify that this thesis which I now submit for examination for the award of Doctor of Philosophy, is entirely my own work and has not been taken from the work of others, save and to the extent that such work has been cited and acknowledged within the text of my work.

This thesis was prepared according to the regulations for postgraduate study by research of the Dublin Institute of Technology and has not been submitted in whole or in part for an award in any other third level institution.

The work reported on in this thesis conforms to the principles and requirements of the DIT's guidelines for ethics in research.

DIT has permission to keep, lend or copy this thesis in whole or in part, on condition that any such use of the material of the thesis be duly acknowledged.

Signature_____ Date_____

Acknowledgements

Over the last few years I've had the pleasure of working with and getting to know some great people. I would like to express my heartfelt gratitude to following people.

First, I would like to express my gratitude to my supervisors, Dr. Brian Mac Namee and Dr. Sarah Jane Delany, for their patience, encouragement, expertise, and trust. Brian was fantastic at communicating his ideas and understanding my mumbled, half-baked explanations. His affable personality and positive outlook transformed a seemingly technical research discipline into an engaging and interesting topic. Sarah Jane was a superb mentor and motivator. Under her tutelage I learnt to develop a focus necessary for scientific research. Their tireless enthusiasm was always exemplified by their draft paper comments, late night/early morning emails, and constant support. I would like to thank my advisory supervisor Prof. Pádraig Cunningham for his support and advice.

To my friends and colleagues at the Applied Intelligence Research Centre (AIRC) in K107A, thank you for your friendship and support. I learnt a lot working with the AIRC group who also provided me with some happy memories. Dr. Rong (Amy) Hu was the office trailblazer. Amy is not just a great scholar, but also a great person. I'd like to thank

my friend Patrick Lindstrom. Paddy was always kind with his time and always offered valuable insights to even the most lowbrow, or trivial discussion. Colm Sloan is a very considerate friend and kept me entertained with some epic yarns and contrarian viewpoints. My thanks also to Niels Schuette for his friendship, tolerance and wit. Yan Li, too, for sharing some interesting and humorous conversations. Dr. Robert Ross and Dr. John Kelleher also deserve mention for their sensible advice.

Prior to undertaking this PhD I knew next to nothing about credit scoring. In this regard, I owe a debt of gratitude to Aoife Darcy (The Analytics Store) who explained the important aspects of credit scoring and statistics. I would also like to thank the Irish Credit Bureau (Michael O'Sullivan, Neil Watson, and Seamus O'Tighearnaigh) for providing me with data and assistance. I should also thank Noel Gilmer (Harland Financial Solutions) and Dr. Christian Thun (Moody's) for their diligent proof reading.

I would like to thank members of the DIT staff who have helped and guided me over the last few years, including Dr. Susan McKeever, Dr. Bryan Duggan, and Dr. Ronan Fitzpatrick. In particular, I offer my thanks to Kevin O'Donnell for having such a positive influence on my formal education.

Without doubt my parents, Patrick and Maureen Kennedy, have made tremendous sacrifices to ensure that their children received a good education and were raised to be honest and to treat others fairly. Thanks also to my siblings, Louise, Padraig, and Cathal, for their gentle ribbing

and support.

I'd like to thank my parent-in-laws, Sean and Ann O'Sullivan, who have placed great faith in me and shown vast amounts of patience and generosity. Thanks also to my sister-in-law, Eavan O'Sullivan, and housemate, Majella Butler. Undoubtedly I have, at times, been completely self-absorbed and ignorant of my surroundings. It would be unfair not to acknowledge the companionship of both Jack Russells, Jack and Bambi. Finally, my wife Muireann who is an extraordinary woman and the bedrock of my life. Her love and support means so much more than any academic award.

Contents

1	Introduction	1
1.1	Background	1
1.1.1	Machine Learning	2
1.1.2	Quantitative Credit Scoring	3
1.2	Motivation	5
1.3	Contributions of the Thesis	8
1.4	Outline of the Thesis	11
1.5	Publications	13
2	Machine Learning and Classification	15
2.1	Basic Concepts	16
2.1.1	The Learning Problem	17
2.1.2	Risk Minimisation	18
2.2	Supervised Classification Algorithms	20
2.2.1	Fisher's Linear Discriminant Analysis (LDA)	20
2.2.2	Logistic Regression	23
2.2.3	Linear Bayes Normal	26
2.2.4	Quadratic Bayes Normal	26
2.2.5	Naïve Bayes Kernel Estimation	26

2.2.6	Support Vector Machines	27
2.2.7	Neural Network, Back Propagation Feed-Forward Network . . .	27
2.2.8	<i>k</i> -Nearest Neighbour	28
2.3	The Class Imbalance Problem	29
2.4	One-Class Classification Techniques	31
2.4.1	Gaussian	34
2.4.2	Mixture of Gaussians	35
2.4.3	Parzen Density Estimation	35
2.4.4	Naïve Parzen	36
2.4.5	<i>k</i> -Nearest Neighbour	37
2.4.6	Support Vector Domain Description	37
2.4.7	<i>k</i> -Means	38
2.4.8	Auto-encoders	38
2.5	Evaluating Classifier Performance	39
2.5.1	Performance Measures	40
2.5.1.1	Confusion Matrix	40
2.5.1.2	Receiver Operating Characteristic Curve	44
2.5.1.3	<i>H</i> Measure	46
2.5.2	Error Estimation	47
2.5.3	Statistical Significance Testing	48
2.6	Conclusion	49
3	Credit Scoring	51
3.1	Background	52

3.1.1	The Basel II Capital Accord	57
3.2	Credit Scorecards	64
3.2.1	Dataset Construction	70
3.2.1.1	Data Quality	70
3.2.1.2	Data Quantity	71
3.2.1.3	Sampling Period	72
3.2.1.4	Class Label Definition	73
3.2.1.5	Dataset Completion	74
3.2.2	Modelling	75
3.2.2.1	Feature Selection	75
3.2.2.2	Coarse Classification	81
3.2.2.3	Reject Inference	85
3.2.2.4	Segmentation	88
3.2.2.5	Model Training	90
3.2.2.6	Scaling	91
3.2.2.7	Validation	94
3.3	Conclusion	95
4	Credit Scoring Challenges	99
4.1	The Low-Default Portfolio Problem	100
4.1.1	Calibration of Low-Default Portfolios	101
4.1.2	Modelling Low-Default Portfolios	103
4.1.3	Low-Default Portfolios: Thesis Research	107
4.2	Behavioural Scoring	107

4.2.1	Behavioural Scoring: Approaches	109
4.2.2	Behavioural Scoring: Thesis Research	112
4.3	Artificial Data	113
4.3.1	Artificial Data: Previous Work	117
4.3.2	Artificial Data: Thesis Research	119
4.4	Conclusion	120
5	Using Semi-supervised Classifiers for Credit Scoring	123
5.1	Evaluation Experiment	124
5.1.1	Data	125
5.1.2	Performance Measures	128
5.1.3	Methodology	129
5.1.4	Classifier Tuning	132
5.2	Results and Discussion	134
5.2.1	Two-class Classifier Performance with Imbalance	138
5.2.2	The Impact of Oversampling	142
5.2.3	One-class Classifiers	147
5.2.4	Optimising the Threshold	148
5.3	Conclusions	156
6	Benchmarking Behavioural Scoring	161
6.1	Experiment Set-up	163
6.1.1	Data	165
6.1.1.1	Dataset Preparation	166
6.1.1.2	Data Generation	170

6.1.1.3	Dataset Labelling	175
6.1.2	Performance Measures	176
6.1.3	Methodology	178
6.1.4	Model Training	180
6.2	Results and Discussion	181
6.2.1	Performance Window Selection	183
6.2.2	Outcome Window Selection	188
6.2.3	Current Status versus Worst Status	193
6.3	Conclusion	196
7	Artificial Data	199
7.1	Methodology	201
7.1.1	Feature Value Generation	201
7.1.2	Label Application	209
7.1.2.1	Label Application: Step 1	210
7.1.2.2	Label Application: Step 2	221
7.1.2.3	Label Application: Step 3	221
7.1.3	Summary	222
7.2	Illustrative Example: Population Drift	223
7.2.1	Framework Configuration	224
7.2.2	Outcome	229
7.3	Conclusions	230
8	Conclusions	233
8.1	Introduction	233

8.2	Summary of Contributions and Achievements	234
8.3	Open Problems and Future Work	238
8.3.1	Low-Default Portfolios	238
8.3.2	Behavioural Scoring	239
8.3.3	Artificial Data	240
A	Notation	241
B	Abbreviations	243
C	Additional Material for Chapter 5	247
D	Additional Material for Chapter 7	297
D.1	Prior Probabilities	297
D.2	Conditional Prior Probabilities	299
D.3	Additional Default Settings	307
References		348

List of Tables

1.1 Contributions, corresponding chapters and publications.	14
2.1 Confusion matrix for binary classification	41
3.1 Application scorecard with a credit score for applicant X	66
3.2 Correlation matrix for bivariate and pairwise correlation. The diagonal of the matrix has values of 1.00 because a variable always has a perfect correlation with itself.	79
3.3 Analysis of a grouped feature. G = goods, B = bads.	84
4.1 Behavioural scoring data sources and associated feature examples (McNab & Wynn, 2000).	110
5.1 Characteristics of the nine datasets used in the evaluation experiment. # Numeric refers to the number of continuous features and # Nominal refers to the number of categorical features.	127

5.2	Test set H measure performance using the normal process on two-class classifiers. The best test set H measure at each class imbalance ratio is underlined. The rank of the different classifiers at each class imbalance ratio is averaged over all the datasets and reported as the AR (average rank). For legibility the H measure figures have been scaled and should be multiplied by 10^{-2} .	141
5.3	Test set H measure performance using the oversample process on two-class classifiers. The best test set H measure for each class imbalance ratio is underlined. The rank of the different classifiers at each class imbalance ratio is averaged over all the datasets and reported as the AR (average rank). For legibility the H measure figures have been scaled and should be multiplied by 10^{-2} .	141
5.4	Average difference in test set H measure performance, oversample process <i>versus</i> normal process. A positive figure indicates the oversample process outperformed the normal process.	146
5.5	Test set H measure performance of logistic regression normal process (LOG_Norm), and OCC process at a class imbalance ratio of 99:1. The best test set H measure per dataset is underlined. The average rank (AR) of the classifiers is also provided. H measure figures should be multiplied by 10^{-2} . $Aus =$ Australia, $Ger =$ German.	148
5.6	Test set harmonic mean performance of Default threshold (D) <i>versus</i> Optimised threshold (O) at a class imbalance ratio 90:10 using the Australia (<i>Aus</i>) and German (<i>Ger</i>) datasets. Harmonic mean figures should be multiplied by 10^{-2} .	149

- 5.7 Test set harmonic mean performance using the normal process on
two-class classifiers. The best test set harmonic mean for each class
imbalance ratio is underlined. The rank of the different classifiers
at each class imbalance ratio is averaged over all the datasets and
reported as the AR (average rank). For legibility the harmonic mean
figures have been scaled and should be multiplied by 10^{-2} 153
- 5.8 Test set harmonic mean performance using the oversample process on
two-class classifiers. The best test set harmonic mean for each class
imbalance ratio is underlined. The rank of the different classifiers
at each class imbalance ratio is averaged over all the datasets and
reported as the AR (average rank). For legibility the harmonic mean
figures have been scaled and should be multiplied by 10^{-2} 153
- 5.9 Average difference in test set harmonic mean performance, oversam-
ple process versus normal process. Positive figure indicates oversam-
ple process outperformed normal process. 154
- 5.10 Test set harmonic mean performance of logistic regression normal
process (LOG_Norm), and OCC process at a class imbalance ratio of
99:1. The best test set harmonic mean per dataset is underlined. The
average rank (AR) of the classifiers is also provided. Harmonic mean
figures should be multiplied by 10^{-2} . *Aus* = Australia, *Ger* = German. 155
- 6.1 ICB data features. Features removed from the ICB data during the
dataset preparation step are indicated by *. 168

6.2 Combination features generated based on the <i>arrears</i> repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.	172
6.3 Combination features generated based on the <i>normal</i> repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.	173
6.4 Combination features generated based on the <i>moratorium</i> (Morat.) repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.	173
6.5 Combination features generated based on the <i>dormant</i> repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.	174
6.6 Combination features generated based on the <i>litigation</i> (Litig.) repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.	174
6.7 Combination features generated based on the <i>frozen</i> repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.	174

6.8	Combination features generated based on the <i>closed</i> repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size. Note: Once Closed, the loan status remains unchanged.	175
6.9	Average class accuracy <i>post hoc</i> analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for the <i>worst status</i> (worst) label definition approach are provided. Note, no statistical significance was detected between the average class accuracies using O24. Statistical significance is indicated by *.	184
6.10	Average class accuracy of LR models trained with data based on a 12-month performance window, <i>post hoc</i> analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for both label definition approaches: <i>worst status</i> (worst) and <i>current status</i> (current) are provided. Statistical significance is indicated by *.	191
7.1	Artificial dataset features	202
7.2	The conditional prior probabilities (Dependency) of each feature with a brief explanation	208
7.3	Loan Rate prior probabilities. The loan rate values used when calculating the monthly loan repayments are also provided.	209
7.4	A list of conditional prior probabilities (CPP) for the FTB feature. Due to a lack of available statistical information we do not differentiate between the CPP of properties located outside of Dublin, the country's main population centre.	210

7.5 Credit Risk Score of a single instance	220
D.1 Location prior probabilities.	297
D.2 New Home prior probabilities.	297
D.3 Loan Rate prior probabilities and loan rate values used when calculating the monthly loan repayments are also provided.	298
D.4 Age group conditional prior probabilities. Each column should total 100%.	299
D.5 LTV conditional prior probabilities. NH = New Home, OH = Old Home, NFTB = Not First-Time-buyer. Each row should total 100%.	299
D.6 First-Time-Buyer (FTB) conditional prior probability (CPP). NFTB = Not First-Time-buyer.	300
D.7 Loan Value conditional prior probabilities. NH = New Home, OH = Old Home, NFTB = Not First-Time-buyer. Each row should total 100%.	301
D.8 Income Group conditional prior probabilities. Each row should total 100%.	302
D.9 Loan Term conditional prior probabilities. Each row should total 100%.	302
D.10 Occupation conditional prior probabilities. M/E = Managerial/Employer. The column of each division should total 100%.	303
D.11 Employment conditional prior probabilities. M/E = Managerial/Employer. Each column should total 100%.	304
D.12 Household conditional prior probabilities. The column of each division should total 100%.	305

D.13 Education conditional prior probabilities.	305
D.14 Expenses-to-Household conditional prior probabilities.	305
D.15 Expenses-to-Income conditional prior probabilities.	306
D.16 Loan Value categories.	308
D.17 House Value categories.	308
D.18 Distribution of the Overall Default Rate across the risk groups. . . .	308
D.19 Risk level scores	309

List of Figures

2.1	Samples from two classes (red and blue) and their histograms resulting from projection onto the line joining the class means. Reproduced from Bishop (2006)	21
2.2	Samples from two classes (red and blue) and their histograms resulting from projection onto the line based on the Fisher's linear discriminant analysis. Reproduced from Bishop (2006)	22
2.3	Example of a classifier's ROC curve, as represented by the blue line. The y -axis represents the TPR and the x -axis represents the FPR. Conversely, the y -axis can represent the FNR and the x -axis the TNR.	44
3.1	Total consumer credit owned and securitised (seasonally adjusted). Source: Federal Reserve Board	55
3.2	Delinquency rate on all real estate loans, all banks, seasonally adjusted. Source: Federal Reserve Board	58
3.3	A process model for developing a credit scorecard, E.C.A. = expert committee approval. Note: Reject inference is performed during application scoring, but not during behavioural scoring. Based on van Gestel & Baesens (2009)	68
3.4	Scorecard scaling using linear scaling.	94

4.1	Behavioural scoring performance window and outcome window.	110
5.1	Normal process; training set - TRAIN, validation set - VALIDATE, test set - TEST.	130
5.2	Oversample process; training set - TRAIN, validation set - VALIDATE, test set - TEST.	131
5.3	Australia: Normal process and one-class classification process test set H measure performance. Selected class imbalance ratios are also highlighted at 70:30, 80:20 and 90:10.	135
5.4	German: Normal process and one-class classification process test set H measure performance.	136
5.5	Thomas: Normal process and one-class classification process test set H measure performance.	137
5.6	Australia: Oversample process and one-class classification process test set H measure performance.	143
5.7	German: Oversample process and one-class classification process test set H measure performance.	144
5.8	Thomas: Oversample process and one-class classification process test set H measure performance.	145
5.9	Australia: Normal process and one-class classification process test set harmonic mean performance.	151
6.1	Behavioural scoring performance window and outcome window.	164
6.2	Example of a customer account detailed by 11 data records.	171

6.3	Monthly default rate, up until December 2010, for customer accounts opened in 2003 and 2004.	176
6.4	Experiment set-up for a 12-month performance window and a 3-month outcome window (Out).	179
6.5	Average class accuracies (<i>y</i> -axis) of the behavioural scoring classification model when a particular combination of performance window and outcome window size definitions are used. The <i>worst status</i> label definition in all cases.	185
6.6	Average class accuracies (<i>y</i> -axis) of the behavioural scoring classification model when a particular combination of performance window and outcome window size definitions are used. The <i>current status</i> label definition in all cases.	186
6.7	Average class accuracy (ACA) comparison of LR models using 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. <i>Worst status</i> label definition. The performance window is fixed at 12-months.	189
6.8	Average class accuracy (ACA) comparison of 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. <i>Current status</i> label definition approach. The performance window is fixed at 12-months.	190
6.9	LR model average class accuracies (<i>y</i> -axis) trained with data based on a 12-month performance window and each outcome window. <i>Current status</i> versus <i>worst status</i>	195

7.1	Data Labelling process.	211
7.2	Example of Coded Rules: Calculation of the risk level for the Location feature for a mortgage applicant instance with a Location value of <i>Galway</i> , a Home Value in the range <i>400k - 500k</i> and a <i>skilled Occupation</i>	214
7.3	Curvilinear transformation: The Location risk level (5) detailed in Figure 7.2 is converted into a Location risk score (17.05) using a user-definable transformation function.	219
7.4	Histogram of Location feature for drift datasets. By way of adjusting the prior probabilities of the Location feature, this scenario simulates gradual decrements to the Dublin, Cork, and Other locations along with gradual increments to the Galway, Limerick, and Waterford locations.	227
7.5	Histogram of FTB feature for drift datasets. The prior probabilities of FTB and non-FTB are gradually reversed.	227
7.6	Histogram of Age Group feature for drift datasets. The conditional prior probabilities of the two attributes representing 26-to-35 year-olds (26-30, 31-35) are reduced in each phase. The conditional prior probability 36-40 Age Group category remains unchanged. The remaining Age Group attributes (18-25, 41-45, 46+) are increased in each of the 5 phases.	228
7.7	Probability density function (PDF) of Credit Risk Score in Phase 1, Phase 3, and Phase 5	228

7.8 The performance of the logistic regression model on the drift and non-drift datasets generated. Performance is measured using the AUC based on a moving average over 3 datasets (where each Phase consists of 3 datasets)	230
C.1 Iran: Normal process and one-class classification process test set H measure performance.	249
C.2 Japan: Normal process and one-class classification process test set H measure performance.	250
C.3 PAKDD: Normal process and one-class classification process test set H measure performance.	251
C.4 Poland: Normal process and one-class classification process test set H measure performance.	252
C.5 Spain: Normal process and one-class classification process test set H measure performance.	253
C.6 UCSD: Normal process and one-class classification process test set H measure performance.	254
C.7 Iran: Oversample process and one-class classification process test set H measure performance.	255
C.8 Japan: Oversample process and one-class classification process test set H measure performance.	256
C.9 PAKDD: Oversample process and one-class classification process test set H measure performance.	257

C.10 Poland: Oversample process and one-class classification process test	
set H measure performance.	258
C.11 Spain: Oversample process and one-class classification process test	
set H measure performance.	259
C.12 UCSD: Oversample process and one-class classification process test	
set H measure performance.	260
C.13 German: Normal process and one-class classification process test set	
harmonic mean performance.	261
C.14 Iran: Normal process and one-class classification process test set har-	
monic mean performance.	262
C.15 Japan: Normal process and one-class classification process test set	
harmonic mean performance.	263
C.16 PAKDD: Normal process and one-class classification process test set	
harmonic mean performance.	264
C.17 Poland: Normal process and one-class classification process test set	
harmonic mean performance.	265
C.18 Spain: Normal process and one-class classification process test set	
harmonic mean performance.	266
C.19 Thomas: Normal process and one-class classification process test set	
harmonic mean performance.	267
C.20 UCSD: Normal process and one-class classification process test set	
harmonic mean performance.	268
C.21 Australia: Oversample process and one-class classification process	
test set harmonic mean performance.	269

C.22 German: Oversample process and one-class classification process test	
set harmonic mean performance.	270
C.23 Iran: Oversample process and one-class classification process test set	
harmonic mean performance.	271
C.24 Japan: Oversample process and one-class classification process test	
set harmonic mean performance.	272
C.25 PAKDD: Oversample process and one-class classification process test	
set harmonic mean performance.	273
C.26 Poland: Oversample process and one-class classification process test	
set harmonic mean performance.	274
C.27 Spain: Oversample process and one-class classification process test	
set harmonic mean performance.	275
C.28 Thomas: Oversample process and one-class classification process test	
set harmonic mean performance.	276
C.29 UCSD: Oversample process and one-class classification process test	
set harmonic mean performance.	277
C.30 Australia: Normal process and one-class classification process test set	
AUC performance.	278
C.31 German: Normal process and one-class classification process test set	
AUC performance.	279
C.32 Iran: Normal process and one-class classification process test set AUC	
performance.	280
C.33 Japan: Normal process and one-class classification process test set	
AUC performance.	281

C.34 PAKDD: Normal process and one-class classification process test set	
AUC performance.	282
C.35 Poland: Normal process and one-class classification process test set	
AUC performance.	283
C.36 Spain: Normal process and one-class classification process test set	
AUC performance.	284
C.37 Thomas: Normal process and one-class classification process test set	
AUC performance.	285
C.38 UCSD: Normal process and one-class classification process test set	
AUC performance.	286
C.39 Australia: Oversample process and one-class classification process	
test set AUC performance.	287
C.40 German: Oversample process and one-class classification process test	
set AUC performance.	288
C.41 Iran: Oversample process and one-class classification process test set	
AUC performance.	289
C.42 Japan: Oversample process and one-class classification process test	
set AUC performance.	290
C.43 PAKDD: Oversample process and one-class classification process test	
set AUC performance.	291
C.44 Poland: Oversample process and one-class classification process test	
set AUC performance.	292
C.45 Spain: Oversample process and one-class classification process test	
set AUC performance.	293

C.46 Thomas: Oversample process and one-class classification process test	
set AUC performance.	294
C.47 UCSD: Oversample process and one-class classification process test	
set AUC performance.	295

CHAPTER

1

Introduction

This chapter introduces the research topic, machine learning and quantitative credit scoring, and its importance to financial institutions. The motivations for the research are discussed along with a brief description of its significance. The aims and contributions of the thesis are then specified. Finally, the chapter concludes with a high-level summary of the organisation of this thesis.

1.1 Background

To position the contribution of this thesis, we begin with a high-level overview of machine learning and quantitative credit scoring.

1.1.1 Machine Learning

Artificial intelligence (AI) (McCarthy *et al.*, 1955) is a field of study that draws from many disciplines including computer science, mathematics and information theory, cognitive psychology, and philosophy (Cook & Holder, 2001). The goal of AI is to develop systems that provide solutions to tasks that have traditionally been regarded as the preserve of intelligent biological systems. As a result of its multi-disciplinary nature, AI-based systems are the manifestation of a broad spectrum of technologies and strategies focused on the development of (Mira, 2008): (i) conceptual models; (ii) the formal representation of these models; and (iii) programming strategies and hardware to implement such models.

A requirement of an AI-based system is the ability to adapt to changes in its environment. Machine learning is a discipline within AI concerned with the programming of computers to automatically adapt and learn from data or past experience (Mitchell, 1997). This can be achieved using an algorithm that specifies a sequence of instructions which transforms the input to output (Alpaydin, 2004). In machine learning, algorithms are used to distinguish between meaningful and irrelevant patterns in data. Examples of machine learning applications include the provision of accurate medical diagnostics (e.g. breast cancer), real-time map-based monitoring of environmental disasters (e.g. forest fires), and sensory monitoring in the industrial process (e.g. mechanical failure).

Supervised learning is a core area of machine learning. In supervised learning the goal is to learn a mapping from the input to the output. The input is data that describes a collection of individual objects of interest and are commonly referred

to as *instances* or *examples*. The output is some outcome or result provided by a supervisor. Classification is a form of supervised learning whereby a mapping (or discriminant function) separates different classes of the instances. The different classes are specified by the output which, in machine learning, is termed as the *class label*. The discriminant function is referred to as a *classifier* or a *model*. A set of instances with their known class label is termed a *training set*. During classification, a model is defined by a set of parameters that are optimised to generate a mapping from training set instances to training set labels. The trained model can be used to *classify* or *label* new, unseen instances.

One-class classification (OCC) is a recognition-based methodology that draws from a single class of examples to identify the normal or expected behaviour of the target class. This is a form of semi-supervised classification as the training data consist of labelled examples for the target class only. This is in contrast to standard supervised classification techniques that use a discrimination-based methodology to distinguish between examples of different classes. OCC techniques have been applied to a wide range of real-world problems such as machine fault detection (Sarmiento *et al.*, 2005), fraud detection (Juszczak *et al.*, 2008), and identity verification (Hempstalk, 2009).

1.1.2 Quantitative Credit Scoring

The term *credit scoring* is used to describe the process of evaluating the risk a customer poses of defaulting on a financial obligation (Hand & Henley, 1997). The objective is to assign customers to one of two groups: *good* and *bad*. A member of the good group is considered likely to repay their financial obligation. A member

of the bad group is considered likely to default on their financial obligation. In its simplest incarnation a credit scorecard consists of a set of characteristics that are used to assign a credit score to a customer indicating their risk level. This credit score can then be compared with a threshold in order to make a lending decision. As credit scoring is essentially a discrimination problem (good or bad), one may resort to the numerous classification techniques that have been suggested in the literature (see Lee & Chen, 2005).

Based on both the task and data used, credit scoring is traditionally divided into two broad types (Bijak & Thomas, 2012). The first, *application scoring*, is used at the time an application for credit is made and estimates an applicant's likelihood of default in a given time period. The data used for model fitting for this task generally consists of financial and demographic information about a sample of previous applicants along with their good/bad status at some later date. The second type of credit scoring, *behavioural scoring*, is used after credit has been granted and estimates an existing customer's likelihood of default in a given time period. Behavioural scoring allows lenders to regularly monitor customers and help coordinate customer-level decision making. The data used for model fitting for this task is based on the customers' loan repayment performance and also their good/bad status at some later date. To be profitable a bank must accurately predict customers' likelihood of default over different time horizons (1 month, 3 months, 6 months, etc.). Customers with a high risk of default can then be flagged allowing the bank to take appropriate action to protect or limit itself from losses.

1.2 Motivation

The upheaval in the financial markets that accompanied the 2007-2008 sub-prime mortgage crisis has emphasised the large proportion of the banking industry based on consumer lending (Thomas, 2009b). Credit scoring is an important part of the consumer lending process. It is an endeavour regarded as one of the most popular application fields for both data mining and operational research techniques (Baesens *et al.*, 2009). Improving the scoring accuracy of the credit decision by as much as a fraction of a percent can result in significant future savings (West, 2000). Furthermore, global (Bank for International Settlements) and national (central banks) regulators insist that financial institutions keep better track of their credit scoring systems. The costs of incorrectly classifying a customer can be high, both financially and in terms of reputation.

The development of a typical credit scorecard can be represented in three main stages, namely (van Gestel & Baesens, 2009, pp.252):

- Dataset Construction
 - In this stage the raw data is collected for preprocessing and preparation (i.e. data cleansing, sampling period, label definition).
- Modelling
 - This stage involves selecting an appropriate classification approach, e.g. linear modelling, neural network architecture, kernel based learning. This stage also includes refining the data using feature selection, feature transformation and coding techniques. The final datasets (or samples) are then

generated for model evaluation and scorecard construction (e.g. scaling and cut-off scores). If applicable, reject inference is also performed during this stage. Reject inference is a method for inferring how rejected applicants would have behaved had credit been granted.

- Documentation
 - In this stage the model design specifications, information technology (IT) specifications, and user manual are written. Details of how the scorecard complies with regulatory specifications are documented. The steps of the dataset construction and modelling process are clearly reported to facilitate the replication of results. The IT infrastructure and links with external entities is also specified. Finally, a guide describing the functioning of the scorecard is written for the end-users.

Although each stage in the credit scoring development process is essential for the delivery of a well-performing scorecard, the *dataset construction*, and *modelling* stages are of particular interest to this research. The *documentation* stage will not be further investigated in this thesis. However, it should be noted that this stage is important in the development and maintenance of an unbiased scorecard that can help lenders make the right decision. In this thesis we examine a number of specific challenges encountered during the dataset construction and modelling stages in the above framework.

During the modelling stage a supervised learning classifier is implemented to discriminate between customers who are labelled as either good or bad. Improved classifier accuracy helps ensure better scorecard performance. However, specific

challenges may arise during the modelling stage. For example, in credit scoring, low-default portfolios are those for which very few customers are labelled as bad. This makes it problematic for financial institutions to estimate a reliable probability of a customer defaulting on a loan. This thesis assesses the performance of machine learning classifiers in credit scoring and their suitability to low-default portfolios.

After credit has been granted, lenders use behavioural scoring to assess the likelihood of default occurring during some specific outcome period. This assessment is based on customers' repayment performance over a given fixed period. Often the outcome period and fixed performance period are arbitrarily selected, causing instability in making predictions. Behavioural scoring has failed to receive the same attention from researchers as application scoring. The bias for application scoring research can be attributed, in part, to the large volume of data required for behavioural scoring studies. Furthermore, the commercial sensitivities associated with such a large pool of customer data often prohibits the publication of work in this area. The task of generating and assessing behavioural scoring datasets during the modelling stage is addressed in this thesis. This is realised using real-world data to generate a collection of behavioural scoring datasets with varying performance and outcome periods. In addition we also examine separate approaches used to label the data.

The credit scoring literature over the last decade has produced numerous studies examining the issues and challenges that occur during the development of credit scorecards. For example, the performance of various models used to construct credit scorecards have been evaluated, (e.g. Baesens *et al.*, 2003; Chen *et al.*, 2011; West, 2000). A concern is that the data used in many of these studies originates from pri-

vate datasets obtained from financial institutions. Due to non-disclosure agreements and commercial sensitivities, obtaining real credit scoring datasets is a problematic and time consuming task. Therefore, part of this work involves the creation of artificial data in order to examine problems in credit scoring.

1.3 Contributions of the Thesis

In this work, we place particular emphasis on understanding the development of quantitative credit scorecards. In addition, we investigate some of the key problems encountered by both practitioners and academics in the fields of credit scoring and machine learning. In the chapters that follow, we: (i) formalise classification techniques and their application to credit scoring; (ii) investigate the application of various approaches for addressing the low-default portfolio problem in credit scoring; (iii) quantify differences between classification models which have been trained using different implementations of a real-world behavioural scoring dataset; and (iv) address the lack of data sharing in credit scoring. To achieve this we explore three main topics in this thesis:

- the applicability of one-class classification to the low-default portfolio problem;
- quantifying differences in model performance based on dataset specifications (i.e. length of the performance and outcome periods) and the label definition approach in behavioural scoring.
- the design and implementation of a framework to generate artificial credit scoring data for application scoring;

We propose different approaches to investigate these issues through the implementation and evaluation of techniques used to assess credit risk. Our main contributions on these initial topics can be summarised in the following:

- **A benchmark of supervised and semi-supervised classification techniques on imbalanced credit scoring datasets (Chapter 5):** We evaluate the performance of a selection of supervised and semi-supervised classification techniques over a number of credit scoring datasets in which the datasets' class labels are unevenly distributed (i.e. class imbalance). We demonstrate that (Kennedy *et al.*, 2012b):
 - adjusting the threshold value on classifier output yields, in many cases, an improvement in classification performance.
 - oversampling produces no overall improvement to the best performing supervised classification algorithms.
 - both supervised and semi-supervised classification techniques merit consideration when dealing with low-default portfolios.
- **Dataset specification for behavioural scoring (Chapter 6):** We perform an empirical evaluation of the contrasting effects of altering the performance period and outcome period using 7-years worth of data from the Irish market. Our results indicate that a 12-month performance period yields an easier prediction task (that is, it gives the highest assurance that the classification will be correct), when compared with other historical payment periods of varying lengths. Our findings show that the performance of a logistic regression

classifier degrades significantly when the outcome window is extended beyond 6-months (Kennedy *et al.*, 2012c).

- **Class label definition approaches for behavioural scoring (Chapter 6):** We consider different approaches to how the concept of default is defined in behavioural scoring. Typically, whether the customer is identified as a default risk or not is set based on either: (i) whether the account is in default at the end of the outcome period; or (ii) at any time during the outcome period. This work investigates both approaches and finds that the latter approach resulted in an easier classification problem (Kennedy *et al.*, 2012c).
- **Artificial data generation framework (Chapter 7):** We develop a framework for generating artificial credit scoring application data and provide illustrative examples of how the framework can be used in practice, with particular focus on a population drift scenario - an especially difficult scenario to investigate using freely available real data (Kennedy *et al.*, 2011).
- **Credit scorecard development literature review (Chapter 3):** We present a comprehensive review of the credit scorecard development process.
- **A review of classification techniques (Chapter 2):** We explain the implementation of a selection of supervised and semi-supervised classification techniques.

Throughout the duration of this research we have publish our results at different conferences and in different journals. A complete list of published work is provided in Section 1.5. Although we place particular emphasis on credit scoring for retail

loans, the techniques and findings described throughout this thesis are applicable to other areas of credit scoring such as corporate lending and sovereign loans.

1.4 Outline of the Thesis

This thesis is organised as follows:

- Chapter 2 provides a high level overview of concepts from supervised learning. Semi-supervised learning is then introduced and discussed. The implementation details of both supervised and semi-supervised classification techniques are described. An overview of class imbalance is provided. Finally, measures used to evaluate classifier performance are introduced and discussed.
- Chapter 3 surveys the literature relating to the development of credit scorecards.
- Chapter 4 continues on from the previous chapter, focusing on literature relating to: (i) low-default portfolios; (ii) behavioural scoring; and (iii) the generation of artificial data.
- Chapter 5 assesses the performance of supervised classification techniques using datasets modified to replicate the low-default portfolio problem. Two techniques used to address class imbalance, oversampling and adjusting the classification threshold, are also evaluated. The applicability of semi-supervised classification techniques to the low-default portfolio problem are also evaluated.

- Chapter 6 evaluates the contrasting effects of altering the performance period and outcome period using 7-years worth of data from the Irish market. We consider different approaches to how the concept of default is defined.
- Chapter 7 proposes a framework that can be used to generate artificial data that simulates credit scoring scenarios.
- Chapter 8 summarises key contributions of this work and highlights opportunities for future research.

1.5 Publications

This thesis is supported by the following publications:

[**Kennedy *et al.* (2010)**] Kennedy, K., Mac Namee, B. & Delany, S.J.: Learning without default: A study of one-class classification and the low-default portfolio problem. In: Proceedings of 20th Irish Conference on Artificial Intelligence and Cognitive Science. (2010) 174–187.

[**Kennedy *et al.* (2011)**] Kennedy, K., Delany, S.J. & Mac Namee, B.: A Framework for Generating Data to Simulate Application Scoring. In: Credit Scoring and Credit Control XII, Conference Proceedings, Credit Research Centre, Business School, University of Edinburgh, CRC. (2011).

[**Kennedy *et al.* (2012b)**] Kennedy, K., Mac Namee, B. & Delany, S.J.: Using Semi-Supervised Classifiers for Credit Scoring. Journal of the Operational Research Society. (2011) doi:10.1057/jors.2011.30.

[**Kennedy *et al.* (2012c)**] Kennedy, K., Mac Namee, B. & Delany, S.J.: A Window of Opportunity: Assessing Behavioural Scoring. Expert Systems with Applications, 40(4), Mar 2013, 1372–1380.

As a summary, the contributions of this work, the corresponding chapters of this thesis and the publications are shown in Table 1.1.

Table 1.1: Contributions, corresponding chapters and publications.

<i>Contribution</i>	<i>Chapter</i>	<i>Publication</i>
Classification overview	Chapter 2	
Credit scorecard development literature review	Chapter 3	
Low-default portfolios	Chapter 5	Kennedy <i>et al.</i> (2012b)
Behavioural scoring	Chapter 6	Kennedy <i>et al.</i> (2012c)
Artificial data framework	Chapter 7	Kennedy <i>et al.</i> (2011)

CHAPTER 2

Machine Learning and Classification

The goal of machine learning is to develop tools and techniques capable of automating time-consuming human activities in an accurate and timely manner. Machine learning-based systems achieve this goal by attempting to discover regularities in a subset of training data which allows for the generation of hypotheses about the data domain as a whole. As a discipline within artificial intelligence, the performance of a machine learning-based system should improve as it acquires experience or data.

The three most prominent machine learning paradigms are: (i) *supervised learning*; (ii) *unsupervised learning*; and (iii) *semi-supervised learning*. In supervised learning tasks the training data is comprised of input data and a corresponding target value. Cases where the target value is one of a finite number of discrete categories, are called *classification* tasks. In unsupervised learning the training data

consists only of input data. *Clustering* is one such example whereby the goal is to discover similar groups or examples within the data. In semi-supervised learning the training data consists of input data and, in some but not all cases, a corresponding target value.

This chapter begins with an overview of the basic concepts of supervised machine learning, with a particular emphasis on binary classification. A selection of supervised binary classification algorithms are detailed in Section 2.2. An overview of the real-world problem of class imbalance is provided in Section 2.3. The class imbalance problem is characterised as one in which the classes are unevenly distributed. This can result in a rarity of examples for a specific class, presenting challenges for supervised binary classification algorithms attempting to discover regularities in such data. As a solution, Section 2.4 introduces a branch of classification, semi-supervised classification. In addition, a number of semi-supervised classification algorithms are examined. Finally, in Section 2.5 methods to evaluate the performance of classification techniques are detailed.

2.1 Basic Concepts

This section provides a formal expression of the elements involved in supervised learning. In a typical supervised learning setting, a training set S of examples $x \in X$ and their associated output value $y \in Y$ is given. X is the set of all possible examples in the input space where $X = \{x_1, \dots, x_i, \dots, x_n\}$. Typically, each example x is described by a vector of *feature values* or *attributes*. Typically, in machine learning texts, a feature can be considered as one of two data types: (i) numeric - the feature

values are real numbers; or (ii) categorical - the feature values are members of a pre-specified, finite set. Statistics texts differ by extending the data types to include: (i) nominal - the feature values are members of an unordered set, e.g. {Renter, Owner, Other}; (ii) ordinal - the feature values are members of an ordered set, e.g. {High, Medium, Low}; (iii) interval - the feature values are measured in fixed and equal units and are members of an ordered set, e.g. temperature in degrees Fahrenheit; and (iv) ratio - the feature values have the properties of an interval data type, but with an absolute zero point (i.e. no negative values) e.g. income-to-expenses. Y is the set of all possible output values in the output space. The training set S is composed of n tuples (or *instances*).

$$S = \{(x_1, y_1), \dots, (x_i, y_i), \dots, (x_n, y_n)\}$$

Importantly, it is assumed that the examples x in S are independently and identically drawn (i.i.d.) from X which is an unknown but fixed joint probability distribution function $P(x, y)$.

2.1.1 The Learning Problem

Using the training set S , the goal of supervised learning is to approximate a function $h : X \rightarrow Y$ which can map an example x_i to its output value y_i . The mapping function is performed using a learning algorithm which is commonly referred to as an *inducer*. A single instance of an inducer for a specific training set is termed a *classifier* (Rokach, 2010). The space of classifiers or functions H is referred to as the *classifier space* or *hypothesis space*.

Depending on the output values in Y , the types of learning problems are com-

monly defined as: (i) *regression learning*, where $Y = \mathbb{R}$; and (ii) *classification learning*, where $Y = C$ such that C constitutes a set of classes where $C = \{c_1, \dots, c_n\}$.

The focus of this thesis is on the latter, classification learning.

It is worth noting that a general multi-class classification learning problem can be decomposed into a collection of binary classification problems (Xu & Chan, 2003). Therefore, this work considers the binary classification problem as the fundamental problem. In a binary classification problem, the two classes can be labelled by 0 and +1 respectively. For example, the type of borrower in credit scoring (i.e. *good* and *bad*), may be represented as $Y = \{0, +1\}$.

2.1.2 Risk Minimisation

To select the optimal classifier from the hypothesis space a *loss function* is used as a quantitative measure of the agreement between the prediction of $h(x)$ and the desired output y . The optimum function h is the minimum expected error (risk),

$$R(h) = L(h(x), y) = \int L(h(x), y) dP(x, y) \quad (2.1)$$

where L denotes a suitably selected loss function. For binary classification the loss function is usually the 0/1 loss, i.e. $L(h(x), y)$ is 0 if $y = h(x)$ and 1 otherwise.

As the underlying probability distribution $P(x, y)$ is unknown, the risk cannot be minimised directly. Instead, a solution that is close to the minimum expected error is inferred from the available training set S . There are two approaches to address this problem, namely, *generative-based* and *discriminative-based* classification (Cunningham *et al.*, 2008). Generative-based approaches learn a model of the joint

probability $P(x, y)$, or $P(y|x)P(x)$, and the required posterior probabilities are then obtained using Bayes theorem

$$P(y|x) = \frac{P(x|y)P(y)}{\sum_j P(x|j)P(j)} \quad (2.2)$$

where the sum in the denominator is taken over all classes. Discriminative-based approaches learn a direct mapping from the input features x to the class labels y , i.e. the posterior $P(y|x)$.

For both approaches an *induction principle* provides a framework with which to estimate the loss function based on the available information in S (Muller *et al.*, 2001). Inductive principles differ in their quantitative interpretation of the optimum classifier, for example, one such difference arises from the encoding a priori knowledge (Cherkassky & Mulier, 2007). Three commonly used inductive principles include *empirical risk minimization* (see Clemençon *et al.*, 2005), *structural risk minimisation* (see Shawe-Taylor *et al.*, 1998), and *regularisation* (see Friedman *et al.*, 2001).

This section has examined two fundamental concepts of supervised learning. The first concept concerns the use of a learning algorithm to induce a classifier capable of mapping an example x_i to its output value y_i . The second concept relates to the use of a loss function to quantitatively measure the predictions of a classifier with the expected output. The following section examines a number of supervised classification learning algorithms.

2.2 Supervised Classification Algorithms

In this section we compare eight well-known supervised classification methods that are suitable for credit scoring and most of which require minimal parameter tuning.

We discuss statistical classifiers (e.g. logistic regression), Bayesian classifiers, k -nearest neighbour classifiers, neural networks and (linear) support vector machines.

Although numerous other classification methods have been presented in the literature, we limit our discussion to a subset of well-known techniques that are suitable for credit scoring and require minimal parameter tuning.

A detailed description of the most popular credit scoring models, linear discriminant analysis and logistic regression, is provided at the beginning of the section. Thereafter, a brief description of the remaining classifiers used in this thesis is provided.

2.2.1 Fisher's Linear Discriminant Analysis (LDA)

This time-honoured approach generates a linear discriminant score based on the linear combination between the input variables which maximises the ratio of variance between the classes to variance within the classes.

Let $y = w_1x_1 + \dots + w_ix_i + \dots + w_nx_n$ be any linear combination of the characteristics $x = (x_1, \dots, x_i, \dots, x_n)$. Adjusting the components of the weight vector $w = (w_1, \dots, w_i, \dots, w_n)$, results in a projection onto one dimension represented as $y = w^T x + w_0$. Classification is achieved by placing a threshold on y , i.e. w_0 , which is the mid-point of the distance between the means. The objective is to select a projection that best separates two groups.

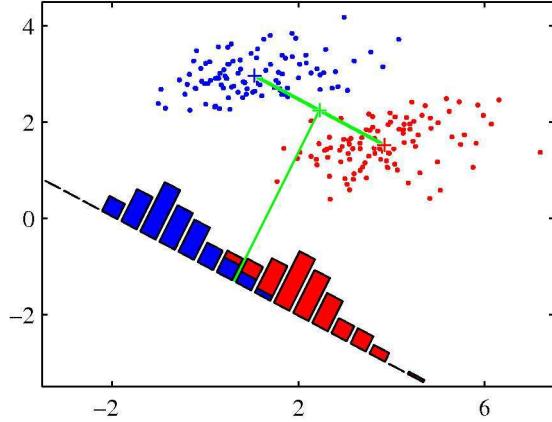


Figure 2.1: Samples from two classes (red and blue) and their histograms resulting from projection onto the line joining the class means. Reproduced from Bishop (2006).

The simplest measure of separation of the classes is the separation of the class means (Bishop, 2006). The weights, w_n , are selected in order to maximize the distance between the means, and w is constrained to have unit length so that $\sum_n w_n = 1$. This approach is illustrated in Figure 2.1 (Bishop, 2006). In the original two-dimensional space both classes (red and blue) are well separated, however a considerable overlap occurs when the samples are projected onto the line joining their means (Bishop, 2006). This problem arises from the strong correlation in the off diagonal matrix of the class distributions, i.e. this approach does not allow for how closely each class clusters together.

Based on the assumption that two groups have a common sample variance, Fisher (1936) suggested a sensible measure of separation as:

$$M = \frac{\text{distance between sample means of two groups}}{(\text{sample variance of each group})^{\frac{1}{2}}} \quad (2.3)$$

where the measure M is the separating distance. This gives a large separation between the class means while also giving a small variance within each class, thereby

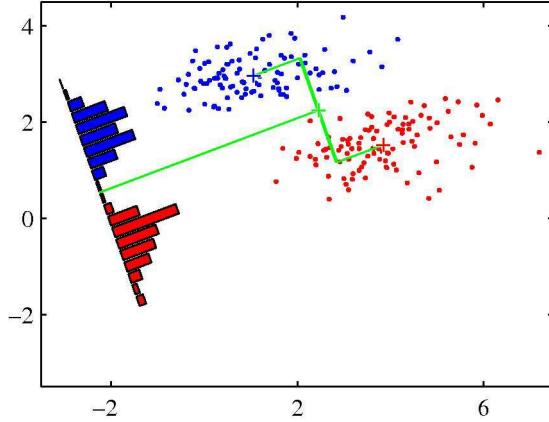


Figure 2.2: Samples from two classes (red and blue) and their histograms resulting from projection onto the line based on the Fisher's linear discriminant analysis. Reproduced from Bishop (2006).

minimizing the class overlap. Assume sample means for the goods and bads are m_g and m_b , respectively, and V is the common sample variance. If $y = w_1x_1 + \dots + w_nx_n$ then the corresponding separating distance M would be:

$$M = w^T \cdot \frac{m_g - m_b}{(w^T \cdot V \cdot w)^{\frac{1}{2}}} \quad (2.4)$$

where w^T is the transpose¹ of w . Differentiating 2.4 with respect to w and setting the derivative equal to zero is maximised when:

$$w^T \propto (V^{-1}(m_g - m_b)^T) \quad (2.5)$$

Figure 2.2 illustrates Fisher's linear discriminant analysis (LDA) and the improved class separation in comparison to Figure 2.1.

An attractive feature of LDA is the fast and simple approach to determine the optimal linear separation, merely requiring simple matrix manipulation such as addition, multiplication, and eigenvalue decomposition (Loog & Duin, 2002). LDA

¹A transpose is performed on a matrix by switching the (i,j) elements with the (j,i) elements.

makes an assumption that the input features are measured on an interval scale or ratio scale. This enables the ranking of objects and a comparison of size differences between them. Unlike other forms of linear discriminant analysis, Fisher's LDA does not require that the input features are independently and randomly sampled from a population having a multivariate normal distribution. LDA assumes that the different groups have equal variance-covariance matrices¹.

2.2.2 Logistic Regression

Logistic regression (see Hosmer & Lemeshow, 2000) is perhaps the most commonly used algorithm within the consumer credit scoring industry (Hand & Zhou, 2009). A regression model outputs a continuous response variable through the linear combinations of predictor variables. As credit scoring is a binary problem we wish to reduce this outcome to 0 or 1. Logistic regression achieves this by applying a logistic transformation, which restricts the output from $[-\infty, +\infty]$ to a probability between 0 and 1. In credit scoring when there are only two outcome groups (i.e. good and bad) binary logistic regression is used. Multinomial logistic regression refers to cases where more than 2 outcome groups are used (i.e. good, indeterminate, bad). Binary logistic regression takes the form of:

$$g(x) = \ln \left(\frac{p_g}{1 - p_g} \right) = b_0 + b_1 x_1 + \dots + b_n x_n \quad (2.6)$$

where p_g is the probability of belonging to the good class. $\frac{p_g}{1 - p_g}$ is called the *odds ratio* and $g(x)$ is the logit transform of p_g . The logit transform is a link function

¹A variance-covariance matrix is a table displaying the variability or spread of the data (variance) and how much two variables move in the same direction (covariance).

used to relate the probabilities of group membership to a linear function of the input features (Worth & Cronin, 2003). Logit has many of the desirable properties of a linear regression model: it is linear in its parameters; may be continuous; and may range from $[-\infty, +\infty]$ depending on the range of x (Al-Ghamdi, 2002). The logit transform is not the only link function available, for example *probit* and *tobit* have been used in credit scoring (see Thomas *et al.*, 2002). However, logit is the easiest to interpret and generally there is little dissimilarity between it and the performance of the probit and tobit link functions (Bewick *et al.*, 2005).

The regression coefficients (b_0 to b_n) are derived using the maximum likelihood estimation (MLE) method (see Kleinbaum & Klein, 2010). The MLE is an iterative and calculation intensive approach which begins by guessing the coefficients values and iteratively changes these values to maximise the log likelihood.

The logistic model produced in Equation 2.6 can be manipulated to estimate the probabilities of class membership (p_g and p_b). The first step is to express the probabilities of class membership in terms of the input features directly:

$$p_g = \frac{\exp(b_0 + b_1x_1 + \dots + b_nx_n)}{(1 + \exp(b_0 + b_1x_1 + \dots + b_nx_n))} \quad (2.7)$$

and the probability of belonging to the bad class:

$$p_b = \frac{1}{(1 + \exp(b_0 + b_1x_1 + \dots + b_nx_n))} \quad (2.8)$$

In the next step, the constant b_0 and the regression coefficients b_1 to b_n are used to define a classification model. According to the following rules an instance can be

defined as belonging to p_g if:

$$b_0 + b_1x_1 + \dots + b_nx_n > 0 \quad (2.9)$$

and, similarly, an instance can be defined as belonging to p_b if:

$$b_0 + b_1x_1 + \dots + b_nx_n < 0 \quad (2.10)$$

If $p_g = p_b$ then an instance has equal probability of belonging to both classes.

These rules are based on a probability cut-off of 0.5. Using a different cut-off value, p_c , the following rules apply:

$$b_0 + b_1x_1 + \dots + b_nx_n > \ln\left(\frac{p_c}{1-p_c}\right) \quad (2.11)$$

and, similarly, an instance can be defined as belonging to p_b if:

$$b_0 + b_1x_1 + \dots + b_nx_n < \ln\left(\frac{p_c}{1-p_c}\right) \quad (2.12)$$

Previously, a disadvantage of logistic regression was the computational intensity required during MLE, however improvements in computer hardware have made this less of an issue. An attraction of logistic regression is that the input features may be either continuous or discrete, or any combination of both types and they do not necessarily have normal distributions (Lee, 2005).

2.2.3 Linear Bayes Normal

This Bayes classifier builds a linear classifier between the classes by assuming normal densities with equal covariance matrices. Classification is performed using the maximum posterior probability rule. See Duda & Hart (1973) for further details.

2.2.4 Quadratic Bayes Normal

This is an extension of linear Bayes normal technique that allows the covariance matrices to be different. Two separate regularisation parameters are used to calculate the covariance matrices. See Duda & Hart (1973) for further details.

2.2.5 Naïve Bayes Kernel Estimation

The naïve Bayes method (see Hand & Yu, 2001) is based on the well established Bayesian approach. It assumes that all attributes are mutually independent of one another given the class label. For each class the method estimates the Gaussian distribution of the attributes. Based on this prior probability, the posterior probability of a previously unseen instance can be determined. Despite its somewhat simplified assumption, naïve Bayes has often proven to be successful, competing with more sophisticated techniques over a variety of applications, particularly in the field of text classification (Bawaneh *et al.*, 2008). Naïve Bayes Kernel Estimation is a generalisation of naïve Bayes which models features using multiple Gaussian distributions. This is known to be more accurate than naïve Bayes which uses a single Gaussian distribution (John & Langley, 1995).

2.2.6 Support Vector Machines

The Support Vector Machine (SVM) approach was first proposed by Vapnik (1995). Using instances from the training data as support vectors, an SVM outlines a class-separating hyperplane in the feature space. Such as to minimise an upper bound of generalisation error, a margin that maximises the distance from the separating hyperplane to the closest support vectors is specified. In order to avoid over-fitting, a soft-margin is used to allow for some misclassification by means of a slack variable. The ‘*kernel trick*’, or Mercers theorem (Mercer, 1909), allows for the data to be mapped to a higher dimensional feature space such that it becomes linearly separable. SVMs have been reported to perform well across a range of domains (van Gestel & Baesens, 2009), including credit risk evaluation. However it is well documented that SVMs are sensitive to parameter selection, difficult to understand due to the lack of transparency, and require a large amount of computation time for large scale classification problems.

2.2.7 Neural Network, Back Propagation Feed-Forward Network

A feed-forward neural network classifier (see Bishop, 1995) consists of a number of processing units (nodes) organised into different layers. The nodes between each layer are interconnected and each connection may have a different weight. The weights on the connections encode the knowledge of the network. Data arrives at the input layer and passes through the network, layer-by-layer, until it arrives at the output layer. There is no feedback between the layers. Each node in a hidden layer

computes a sum based on its input from the previous layer. A sigmoid function condenses the sum into a manageable range and is finally passed to the output layer to produce the final network result. During training, by comparing the differences between the desired output values and actual output values, the connection weights are modified so as to learn the function in question.

2.2.8 k -Nearest Neighbour

The nearest neighbour classifier assigns an instance based on the class of its nearest neighbours. It is more commonly referred to as k -nearest neighbour (k -NN) as it is often more beneficial to consider more than one neighbour (see Henley & Hand, 1996).

This section described eight well-known supervised classification methods that are commonly deployed by machine learning-based systems to automate tasks such as granting credit or detecting fraud. Indeed, we should regard the routine use of such systems in industry, education, and elsewhere as the *ultimate test* for machine learning (Langley & Simon, 1995). It is not uncommon in a real-world setting for observations of a particular class to occur a lot less frequently as compared with normal populations. Quite often the cost of incorrectly classifying samples of a rare class is greater than the contrary case, e.g. classification of a fraudulent transaction as a legal transaction (false negative). In machine learning, a dataset with uneven class distributions can be considered as an imbalanced dataset. The next section identifies the challenges posed by imbalanced datasets to supervised classification.

2.3 The Class Imbalance Problem

Class imbalance presents a problem to most classification algorithms as they assume a balanced distribution of the classes (Japkowicz, 2000). Typically these algorithms attempt to maximise the overall classification accuracy by predicting the most common class (Drummond & Holte, 2005b). That is, they construct models which minimise the number of classification errors, without regard for the significance of misclassifying examples of each class (Seiffert *et al.*, 2009). Whilst such models may indeed be accurate, frequently, particularly for real-world problems, they are not very useful.

For sometime now, the class imbalance problem has attracted a lot of attention within the machine learning community (Bellotti & Crook, 2008; Chawla *et al.*, 2004). Weiss (2004) differentiates between two types of class imbalance (or data rarity) based on whether the rarity is an *absolute* or *relative* property of the data. With absolute rarity the number of examples associated with the minority class is small in an absolute sense. For relative rarity the examples are not rare in an absolute sense but are rare relative to other objects. Both forms of rarity present challenges for conventional machine learning-based systems. Typically where absolute rarity occurs, then there will also be a problem with relative rarity - provided that the dataset is not too small (Weiss, 2004). Furthermore, Weiss (2004) argues that as both problems share many similar characteristics, they also use many of the same solutions.

A number of solutions have been proposed and can be categorised at the data-level and algorithmic-level (Chawla *et al.*, 2004). At the data-level, solutions attempt

to balance the class distribution by resampling the data. Such solutions include, undersampling the majority class and/or oversampling the minority class to ensure that the class distributions are approximately equal. It should be noted that undersampling is an inappropriate solution to the problem of absolute rarity. Random oversampling of the minority class is one of the most common techniques used. One of the drawbacks to using oversampling is the increased possibility of over-fitting since exact copies of rare objects are made which add no new information to the dataset. Over-fitting occurs when the classification algorithm induces a model that is so customised to the training data that it performs poorly on unseen test data (see Hawkins *et al.*, 2004).

At the algorithmic-level, solutions include adjusting the costs of the various classes so as to counter the class imbalance, adjusting the threshold (or cut-off) value on the output of the constructed model, and recognition-based approaches that learn from one-class rather than discrimination-based based approaches (Chawla *et al.*, 2004). The simplest approach to overcome the imbalance problem is to simply adjust the threshold of the constructed model, indeed Provost (2000) cautions that using such models without adjusting the output threshold may well be a “*critical mistake*”. Weiss (2004) provides an overview of the various approaches to dealing with class imbalance.

In this work we are concerned with studying the impact of absolute rarity on the predictive performance of classification models. Along with the problems experienced by class imbalance, rare objects are, by their very nature, atypical and require special attention. One-class classification (OCC) is one such method used to address the data rarity problem.

2.4 One-Class Classification Techniques

Semi-supervised classification, which attempts to learn from both labelled and unlabelled data has attracted much attention in recent years (see Chapelle *et al.*, 2006; Zhu & Goldberg, 2009). The basic premise of semi-supervised classification is to combine unlabelled training data with the labelled data in order to modify and refine the hypothesis for improved classification performance (Cao & He, 2008). For example, self-training semi-supervised learning algorithms (see Wang *et al.*, 2006) induce an initial classifier from the labelled examples in the training data. Using this classifier, the unlabelled examples in the training data are then assigned probabilistic labels and those with the highest probability are appended to the labelled training data. This process is repeated and each time the most confident unlabelled examples are labelled (Cao & He, 2008).

OCC techniques are a form of semi-supervised classification that distinguish a set of target objects from all other objects (Moya *et al.*, 1993). In OCC the training data consists of labelled examples for the target class only, as non-target class examples are too expensive to acquire or too rare to characterise. As OCC techniques do not require labelled examples of the missing class during the initial induction process, it has been described as an *extreme* version of semi-supervised learning (Chawla *et al.*, 2004).

OCC techniques have been successfully applied to a wide range of real-world problems such as:

- fault detection in semi-conductors where it is difficult to collect data under

faulty conditions and even harder to collect data for all possible types and combinations of faults (Sarmiento *et al.*, 2005);

- fraud detection in plastic card transactions. Unlike traditional binary classification techniques, OCC techniques proved adept at identifying new types of fraud as fraudsters change tactics adaptively (Juszczak *et al.*, 2008); and
- identity verification based on continuous typist recognition where only the keystroke patterns of authorised users are known (Hempstalk, 2009).

The term OCC is believed to have originated from Moya *et al.* (1993) and is only one of a number of terms used to describe similar approaches - other terms include *outlier detection* (Ritter & Gallegos, 1997), *novelty detection* (Bishop, 1994), and *concept learning* (Japkowicz, 1999).

Following the taxonomy described by Tax (2001), OCC techniques can be divided into three groups: *density methods*, *boundary methods*, and *reconstruction methods*. This is by no means an exhaustive discrimination, but conceptually it is the simplest and most popular. For a detailed description of OCC taxonomies refer to (Chandola *et al.*, 2009).

All OCC methods share two common elements: a measure of the proximity of an object, z , to the target data; and a threshold, θ , to which the proximity measure is compared. An object, z , is considered to be a member of the target class when the proximity of z to the target data is less than the threshold θ . In this work, we consider the large majority of good payers (i.e. those customers without a loan default) as the target class.

Density estimation approaches to OCC directly estimate the probability distri-

butions of features for the target class by fitting a statistical distribution, such as Gaussian, to the target data. The success of this approach depends on factors such as the target data sample size and whether the selected statistical distribution is appropriate for the target data. The density techniques provide the most complete description of the target data, but as a drawback to this they may require large amounts of data (Tax & Duin, 1999).

OCC approaches based on boundary estimation fit a boundary around the target class data, whilst simultaneously attempting to minimise the volume of the enclosed area. Boundary methods offer a degree of flexibility in that an estimate of the complete probability density is not necessary. The computation of the boundary is based on the distances between the objects in the target data. In some cases a kernel function is used to define a flexible boundary. This approach works well with small sample sizes and an uncharacteristic training dataset (Tax, 2001).

Reconstruction methods are trained to reproduce an input pattern by assuming a model of the data generation process. The parameters of the assumed data generation model are estimated during the learning phase. This differs from density and boundary methods as reconstruction methods do not rely on statistical assumptions made about the data. A reconstruction error is used to determine if the object belongs to the target or outlier class.

A good OCC model should maximise both the number of target objects accepted and outlier objects rejected. Specifying the trade-off between the fraction of target objects accepted and the fraction of outlier objects rejected, through the threshold θ is the most important feature of OCC (Tax, 2001). The threshold is usually adjusted heuristically (and evaluated using a test dataset) to attain the desired trade-off. Too

small a value for θ will cause the model to underfit the data and cover the entire feature space, whereas a large θ will over-fit the data, resulting in a minimised target space.

In some circumstances certain OCC models can incorporate outlier data. The performance of a OCC model may be compromised if outlier data is used as the performance of the model becomes dependent on the outlier data and poor quality data or low quantities of outlier data which are not representative of the problem will damage performance (Hempstalk, 2009). For clarity such models are not used in this study although they may be considered in future work. The remainder of this section will describe each of the OCC algorithms used in this thesis (in all cases the actual implementation used is from the Matlab Data Description Toolbox (DDTools) (Tax, 2009) and some specific details given stem from this).

2.4.1 Gaussian

A density estimation method that assumes the target data is generated from a unimodal multivariate normal distribution, the Gaussian model is one of the simplest OCC techniques (see Tax, 2001). For an object, z , the Mahalanobis distance to the training set distribution is calculated as follows:

$$f(z) = (z - \mu)^T \Sigma^{-1} (z - \mu) \quad (2.13)$$

where μ is the mean and Σ is the covariance matrix of the training set, both of which are estimated using an Expectation-Maximisation (EM) approach. This distance is compared to a threshold θ to make a classification. The Mahalanobis distance is

used in order to avoid numerical instabilities. A caution to the use of the Gaussian method is that if the assumption that the data fits a normal distribution is violated the model may introduce a large bias (Tax, 2001).

2.4.2 Mixture of Gaussians

A mixture of Gaussians model (see Bishop, 1995) is a linear combination of k Gaussian distributions. Although this is a more flexible approach than the single Gaussian method it requires more data as it may display greater variance when only a limited amount of data is available. To build a mixture of gaussians model the training data is divided into k clusters, each of which is modelled by a Gaussian distribution. For an object z a superposition of k Gaussian densities can be written as:

$$f(z) = \sum_{i=1}^k \alpha_i \exp \left\{ -(z - \mu_i)^T \Sigma_i^{-1} (z - \mu_i) \right\} \quad (2.14)$$

where α_i are the mixing coefficients, again μ is the mean and Σ is the covariance matrix. For each cluster i , α_i , μ_i and Σ_i are estimated using the EM algorithm. Given a mixture, the threshold, θ on the density determines if z is classified as target or non-target data.

2.4.3 Parzen Density Estimation

The Parzen density estimator (Parzen, 1962) is an extension of the mixture of Gaussians method. It is a non-parametric technique that uses a kernel to estimate the probability density function. Each object in the target class is treated as the centre of a Gaussian distribution. Based on this, a measure of the likelihood that an object

belongs to the target data is computed by averaging the probability of membership of the Gaussian distributions. Classification is obtained by comparison to a threshold, θ . Let $p(z)$ be the density function to be estimated. Given a set $D = \{z_1, z_2 \dots z_n\}$ of n target objects, the Parzen density estimate of $p(z)$ is:

$$p(z) = \frac{1}{nh} \sum_{i=1}^n \rho \left(\frac{z - z_i}{h} \right) \quad (2.15)$$

where h is a smoothing parameter, and ρ is typically a Gaussian kernel function:

$$\rho(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2} \quad (2.16)$$

The width of the Gaussian kernel, h , is optimised by maximising the likelihood in a leave-one-out fashion, as per Duin (1976).

When large differences in density exist, the Parzen kernel method will give poor results in low density areas. Like all density approaches, it requires a large amount of target data to make a reliable probability density estimation.

2.4.4 Naïve Parzen

The naïve Parzen is a simplification of the Parzen density estimator inspired by the naïve Bayes approach (see Friedman *et al.*, 2001). A Parzen density is estimated in each feature dimension separately, and the probabilities are multiplied to give the final target probability.

2.4.5 k -Nearest Neighbour

The k -nearest neighbour (Cover & Hart, 1967) method can be adopted to construct a one-class classifier. The one-class k -NN (Tax & Duin, 2000) classifier is a boundary-based approach that is based on the number of target objects in a region of a certain volume. Classification is performed using a threshold on the ratio between two distances. The first is the distance between the test object z and its k th nearest neighbour in the training set, $NN(z_i, k)$ (k is a parameter of the approach). The second distance is measured as the distance between the k th nearest training object and its k th nearest neighbour. If the first distance is much larger than the second distance, the object is regarded as a non-target object (Tax & Duin, 2000). The ratio is calculated as follows:

$$p(z) = \frac{\|(z, NN(z, k))\|}{\|(NN(z, k), NN(NN(z, k), k))\|} \quad (2.17)$$

where Euclidean distance is used to measure the distance between objects. For further details refer to Tax (2001).

2.4.6 Support Vector Domain Description

The Support Vector Domain Description (SVDD) (Tax & Duin, 1999) is a kernel-based boundary method which attempts to find the most compact hypersphere that encloses as many target instances as possible. By minimising the volume of the hypersphere, the chance of accepting outlier objects is reduced. To generate a flexible boundary the input space can be mapped into a higher dimensional and

more separable feature space. This transformation is typically performed using a Gaussian kernel. Classification is performed by comparing the distance between an object, z , and the target boundary to a threshold, θ .

2.4.7 k -Means

k -Means clustering (see Bishop, 1995) can be adapted into a relatively straight-forward reconstruction approach to OCC. The approach subdivides the output space, onto which new objects are projected, into k cluster prototypes or centres. The prototypes are located such that the average distance to a prototype centre is minimised as follows:

$$\epsilon_{k-means} = \sum_i (\min_k \|z_i - \mu_k\|^2) \quad (2.18)$$

where μ_k represents the k -th cluster centre. The objects in the training set are clustered, and when a new object is to be classified its distance from the nearest prototype is used as a measure that can be thresholded in order to identify outliers. If the distance is greater than a threshold, θ , the object will be classed as non-target data. A drawback can be that outliers form in clusters by themselves.

2.4.8 Auto-encoders

An *auto-encoder*, also referred to as an *auto-associator*, is a reconstruction based approach introduced by Japkowicz (1999) based on the work of Hinton (1989). An auto-encoder is a particular type of neural network with a single hidden-layer, which is trained to reproduce an input pattern x at the output of the network, $NN(x)$.

Because the network has a narrow hidden layer (or *bottleneck*), it compresses redundancies in the input. This feature can be utilised to train the network to reconstruct examples from a target class as accurately as possible. Such a network will then perform poorly at reconstructing non-target data which present different structural irregularities. Classification is achieved by comparing the reconstruction error when test examples are presented to the network to a threshold.

2.5 Evaluating Classifier Performance

Due to the potentially high costs associated with suboptimal classifier performance it is necessary to evaluate performance both in absolute real terms and relative to other classifiers. The four main components used to evaluate classifier performance are (Japkowicz & Shah, 2011): (i) performance measures; (ii) error estimation; (iii) statistical significance testing; and (iv) test benchmark selection.

The *performance measures* component relates to the selection of metrics used to measure predicted classifier outcomes against the actual outcomes. Once a suitable performance measure has been selected, the *error estimation* component is used to identify an appropriate technique for testing classifier performance. Such techniques attempt to ensure a representative sample of the population is selected with which to assess classifier performance. Doing so provides as unbiased an estimate of the selected performance measure as possible. The *statistical significance testing* component concerns methods used to obtain a precise assessment of the significance of the results measuring classifier performance. The use of statistical methods to justify the selection of a particular classifier is an important field of study in machine learn-

ing (see Dietterich, 1998). The final component, *test benchmark selection*, considers the appropriateness of the selected datasets and domains used to evaluate classifier performance. This is of particular importance when assessing the performance of a learning algorithm over multiple domains. The datasets from each domain may differ considerably in terms of complexity and size, thus distorting the suitability of a particular learning algorithm to a specific domain. As this thesis is concerned with only the credit scoring domain, we do not provide a discussion of the test benchmark selection component. However, a full overview of the issues affecting credit scoring data is provided in Section 4.3. The remainder of this section provides an overview of the first three components, with particular emphasis placed on methods and measures relevant to this thesis.

2.5.1 Performance Measures

This section describes metrics that are often used to evaluate the performance of a classifier. In this thesis binary classifier output is represented as: 1 for accepting (*non-defaulter* or *good*) or 0 for rejecting (*defaulter* or *bad*) a credit applicant. Many ranking classifiers produce a numeric score which can be binarised by the use of a threshold. This section begins with what a confusion matrix is and then continues by describing performance measures derived from the confusion matrix.

2.5.1.1 Confusion Matrix

The decision made by a classifier can be represented in a structure known as a confusion matrix (or contingency table). For binary classification, the confusion matrix is a 2 x 2 matrix with the two classes, commonly referred to as the *positive*

and *negative* class. The confusion matrix has four categories:

- true positive (TP) are *positive* instances correctly classified as *positive*;
- false negative (FN) corresponds to instances classified as *negative* but are actually *positive*;
- true negative (TN) refers to *negative* instances correctly classified as *negative*;
and
- false positive (FP) are *negative* instances incorrectly classified as *positive*.

Table 2.1 displays how a confusion matrix can be presented.

Table 2.1: Confusion matrix for binary classification

TP	FN	A_p
FP	TN	A_n
P_p	P_n	

The acronyms in Table 2.1 are:

- A_p : actual positive class total (TP + FN)
- A_n : actual negative class total (TN + FP)
- P_p : predicted positive class total (TP + FP)
- P_n : predicted negative class total (TN + FN)

Given the numbers from the confusion matrix, several performance measures can be calculated, such as *sensitivity* (Equation 2.19), *specificity* (Equation 2.20), the *false positive rate* (Equation 2.21), the *false negative rate* (Equation 2.22), the *class accuracy* (Equation 2.23), the *average class accuracy* (Equation 2.24), and the

harmonic mean (Equation 2.25). Sensitivity, also known as the *true positive rate* (TPR), measures the proportion of positive (non-default) examples that are predicted to be positive. Specificity, also known as *true negative rate* (TNR), measures the proportion of negative (default) examples that are predicted to be negative. The false positive rate (FPR) measures the proportion of negative examples that are misclassified as positive. The false negative rate (FNR) measures the proportion of positive examples that are misclassified as negative. Class accuracy measures the fraction of correctly classified examples. The *harmonic mean* provides a suitable composite measure of sensitivity and specificity and is calculated as shown in Equation 2.25.

$$\text{Sensitivity} = \frac{TP}{TP + FN} \quad (2.19)$$

$$\text{Specificity} = \frac{TN}{TN + FP} \quad (2.20)$$

$$\text{FPR} = 1 - \text{Specificity} = \frac{FP}{FP + TN} \quad (2.21)$$

$$\text{FNR} = 1 - \text{Sensitivity} = \frac{FN}{FN + TP} \quad (2.22)$$

$$\text{Class Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \quad (2.23)$$

$$\text{Average Class Accuracy} = \frac{\text{TPR} + \text{TNR}}{2} = \frac{\text{Sensitivity} + \text{Specificity}}{2} \quad (2.24)$$

$$\text{Harmonic Mean} = \frac{2 * \text{Sensitivity} * \text{Specificity}}{\text{Sensitivity} + \text{Specificity}} \quad (2.25)$$

Some studies, (e.g. Lessmann *et al.*, 2008), refrain from selecting a classification threshold contending that studies comparing the same classifiers and datasets could easily come to different conclusions as a result of employing different methods for determining classification thresholds. In this thesis we address this issue by clearly defining the average class accuracy and harmonic mean. Both measures assume equal misclassification costs for both false positive and false negative predictions. This may be a problem if we consider that one type of classification error may be a lot more costly than the other. However, in the absence of available cost matrices the average class accuracy and harmonic mean are the most appropriate performance criteria as a means of assessing the accuracy of a classifier at a specific threshold.

The confusion matrix measures classifier performance at a specific threshold, by way of extension, graphical analysis methods and their associated performance measures are used to measure classifier performance over a range of threshold values. Two such methods, the receiver operating characteristic curve and the H measure, are discussed below.

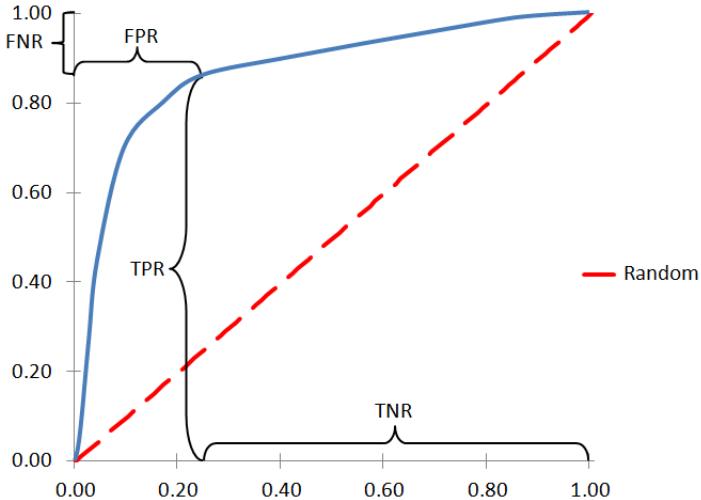


Figure 2.3: Example of a classifier’s ROC curve, as represented by the blue line. The y -axis represents the TPR and the x -axis represents the FPR. Conversely, the y -axis can represent the FNR and the x -axis the TNR.

2.5.1.2 Receiver Operating Characteristic Curve

A receiver operating characteristic (ROC) curve is a plot which displays how the number of correctly classified positive instances varies with the number of incorrectly classified negative instances. Figure 2.3 illustrates the false positive rate on the x -axis against the true positive rate on the y -axis. Each point on the ROC curve represents a classification threshold $\theta \in [0, 1]$ that corresponds to particular values of the false positive rate, and true positive rate. In Figure 2.3 an *operating point* has been selected from which a confusion matrix and associated performance measures can then be calculated.

The ROC space is a unit square as it holds that $0 \leq \text{TPR} \leq 1$ and $0 \leq \text{FPR} \leq 1$. The point $(0,0)$ represents a trivial classifier that classifies all instances as negative. Likewise, the point $(1,1)$ classifies all instances as positive. The diagonal connecting both these points $[(0,0),(1,1)]$ has $\text{TPR} = \text{FPR}$. Any classifier falling along this diagonal line is considered a random classifier as they randomly classify

instances as positive and negative. The point (0,1) represents the perfect classifier as it correctly classifies all instances.

A ROC curve is, essentially, a compilation of confusion matrices over the varying classification thresholds of a classifier. The finite number of instances in a dataset imposes an upper bound on the number of points used to plot the ROC curve. A step function at each point in the ROC space is obtained by varying the classification threshold results. The ROC curve is then plotted by extrapolation over this set of finite points.

To compare the ROC curves of different classifiers, one often calculates the summary statistic *area under the ROC curve* (abbreviated as AUC) (Bradley, 1997; Hanley & McNeil, 1982). The AUC of a classifier is equivalent to the probability that the classifier will rank a randomly chosen positive instance higher than a randomly chosen negative instance (Fawcett, 2006). In Figure 2.3 the AUC is represented by the area beneath the blue ROC curve. For perfect classification this value will be 1, for a classifier which has purely random classification the AUC is 0.5.

A commonly used measure in credit scoring is the Gini coefficient (see Hand, 2005) which corresponds to twice the area between the ROC curve and the diagonal, or:

$$\text{GINI} = 2 * \text{AUC} - 1 \quad (2.26)$$

The AUC and Gini measure performance over all classification thresholds, this may be misleading when we are interested in classifier performance over a narrow range of classification thresholds. Using the ROC plots would give more information, however ROC curves may overlap leading to confusion when selecting the most

suitable classifier.

Finally, another evaluation measured frequently used in credit scoring is the Kolmogorov–Smirnov (KS) statistic, which provides a single value that ranges from 0 to 1. The KS statistic measures the maximum difference between the cumulative distribution functions of the predicted probabilities of the good class and the bad class (Seliya *et al.*, 2009).

2.5.1.3 *H* Measure

The Gini coefficient and the AUC are commonly used in credit scoring to estimate the performance of classification algorithms in the absence of information on the cost of different error types. Hand (2009), however, demonstrates how these measures actually use costs derived from the data used and suggests that their application may produce misleading results about classification performance. For example, the AUC uses a probability distribution of the likely cost values that depend on the actual score distributions of the classifier. As a result the probability distribution of the likely cost values will vary from classifier to classifier, as per the score distribution. This prevents different classifiers from being compared in an equal manner.

As an alternative, Hand proposes the *H* measure (Hand, 2009) that uses a probability distribution of the likely cost values that is independent of the data. This *Beta* distribution (see Hand, 2009) contains two parameters, α and β , that can increase the probability on certain ranges of the cost believed to be more likely. It is recommended that for situations when nothing is known about the costs then a Beta distribution with $\alpha = 2$ and $\beta = 2$ should be used. Normally, *H* measure values range from zero for models which randomly assign class labels, to one for models

which obtain perfect classification. In this thesis we adopt the recommended α and β settings so as to allow for universally comparable results.

Cost curves [for a detailed discussion see Drummond & Holte (2000, 2006)] are a graphical technique for visualising classifier performance over a range of class distributions. Flach *et al.* (2011) argue that the H measure is a linear transformation of the area under the cost curve, with a number of minor variations, e.g. using a Beta distribution instead of the uniform distribution for costs. These minor variations appear to be “*no more strongly justified*” than the area under the cost curve.

2.5.2 Error Estimation

After an appropriate performance measure is selected, the next step is to test the classification algorithm. If the data used to test a classification algorithm is not representative of the actual distribution then any experimental results may lead to unwarranted and unverified conclusions. The purpose of an error estimation technique is to generate as unbiased an estimate of the chosen performance measure as possible. Broadly speaking, error estimation techniques can be split into three separate approaches (Japkowicz & Shah, 2011): (i) resubstitution; (ii) hold-out; and (iii) resampling.

The *resubstitution* error estimate is obtained by using the same dataset to construct a classifier and also to assess its performance (Kim, 2009). This approach is feasible when the entire population or a highly representative sample thereof is available. Thus, leading to a convergence towards the true error rate, as the number of instances used to construct and test the classifier increases.

Using the *hold-out* approach one randomly splits the dataset into a training set

and test set. As its input, the learning algorithm uses the labelled instances of the training set and outputs a classifier. The classifier is then presented with the unlabelled instances of the withheld test set. The classifier predicts the labels for each of these instances and the estimate of the error rate is obtained. Typically, this procedure is repeated many times, with the average of the estimated error rate called the *repeated hold-out estimate*. Often a portion of the training set is set aside for tuning the parameters of the classifier. An advantage of the hold-out approach is the independence of the test set from the training set.

Resampling error estimation methods are used when it is necessary to obtain a sufficiently large enough dataset capable of supporting reliable error estimates. One such frequently used method is *k-fold cross validation* (CV), which partitions the dataset into k-subsets of equal size. At each turn, as per the hold-out approach, one set is used for testing and the remainder for training the learning algorithm. The error estimation is averaged over every partition made.

2.5.3 Statistical Significance Testing

Statistical significance testing is used in scientific research to determine whether the performance difference between classification algorithms are attributable to real factors or apparent factors which are caused by uncontrolled variability of any sort. For example, Brown & Mues (2012) performed a comparison of classification algorithms for imbalanced credit scoring datasets by using a set of designed significance tests. Demšar (2006) recommends a set of non-parametric statistical tests that can be used for comparing the performance of two or more of classifiers over multiple datasets. Namely, the Wilcoxon signed-rank test (Wilcoxon, 1945) when comparing two clas-

sifiers and the Friedman test (Friedman, 1937) when more than two classifiers are compared over multiple datasets. Extensions to Demšar (2006) guidelines and recommendations are provided by García & Herrera (2008) and García *et al.* (2010), including advanced alternatives to the Friedman test. Rather than have classifier performance ranked separately for each dataset, the Friedman aligned ranks test (Hodges & Lehmann, 1962) [or its alternative the Kruskal-Wallis one-way analysis of variance by ranks test (Kruskal & Wallis, 1952)] compare classifier performance among all the datasets of interest.

2.6 Conclusion

In this chapter we outlined the principle concepts and approaches to classification problems. Classification is an important research area in the field of machine learning. The goal of classification is to assign class labels to a set of objects described by a collection of features. In supervised learning, classification is performed by constructing a model using a training dataset of labelled examples. Numerous supervised learning algorithms exist with which to perform classification. These techniques span from well established methods, such as logistic regression and linear discriminant analysis, up to more recent approaches, such as SVMs. Other popular approaches that appear in the credit scoring literature include linear and quadratic Bayes normal, naïve Bayes, neural networks, and k -nearest neighbour.

Most supervised classification algorithms assume a balanced distribution of class labels from which they attempt to maximise the overall class accuracy during the training process. When the class distributions are imbalanced such algorithms often

misclassify examples of the minority class. A number of solutions, categorised at the data-level and algorithmic-level, have been proposed. Adjusting the threshold on classifier output is perhaps the simplest. Another approach is to balance the dataset by resampling the data, e.g. oversampling. For the more extreme cases of class imbalance (i.e. data rarity), a form of semi-supervised learning, one-class classification, is recommended.

Although semi-supervised learning techniques normally use both labelled and unlabelled data during the training process, OCC uses only labelled examples. OCC techniques have been successfully applied to a number of classification problems such as fault detection, fraud detection, and identity verification. All OCC methods share two common elements: a measure of the proximity of an object, z , to the target data; and a threshold, θ , to which the proximity measure is compared. A collection of OCC algorithms were described, ranging from the relatively straight forward Gaussian approach to the computationally intensive SVDD technique.

Finally, measures to evaluate the performance of classification techniques were detailed. These include measures derived from the confusion matrix, AUC, and the H measure. Approaches used to generate an unbiased estimate of the selected performance measure, along with statistical tests to determine the significance of classifier performance were also detailed.

The next chapter gives a literature review on credit scorecards.

CHAPTER 3

Credit Scoring

As discussed in Chapter 2, the goal of classification is to correctly assign class labels to previously unseen instances of a dataset. For financial institutions, classification systems, in the form of credit scoring, are used on a daily basis to assess the credit risks associated with lending to a customer. In credit scoring, a customer's *creditworthiness* describes their ability and willingness to repay a financial obligation (e.g. a loan). *Credit risk* is defined as the risk of loss arising due to any real or perceived change in a customer's creditworthiness (Anderson, 2007). In this thesis we use the term *credit scoring* to describe the set of decision models and techniques used by lenders to rank and assess the credit risk presented by different customers. The objective of credit scoring is to assign both existing and prospective customers to one of two groups: *good* or *bad*. A member of the *good* group is considered likely to repay their financial obligation. A member of the *bad* group is considered likely to default on their financial obligation. Generally, credit scoring models are categorised into

two different types, *application scoring* and *behavioural scoring*. Application scoring attempts to predict a customer's default risk at the time an application for credit is made, based on information such as applicant demographics and credit bureau records. Behavioural scoring assesses the risk of existing customers based on their recent accounting transactions.

This chapter examines credit scoring for loans to households and individuals (i.e. retail loans) and describes the credit scoring development process, previously introduced in Section 1.2. The rest of this chapter is organised as follows. Section 3.1 presents a brief overview of the requirement for credit scoring. Section 3.2 begins with a description of credit scorecards (a decision making tool used to accomplish credit scoring) and then describes the development process used to construct a credit scorecard. The two main stages in the development process are (i) dataset construction and (ii) modelling which are described in Section 3.2.1 and Section 3.2.2, respectively. Finally, a conclusion is provided in Section 3.3.

Other good, detailed discussions on credit scoring can be found in research articles by Thomas *et al.* (2005), Crook *et al.* (2007), Thomas (2009b), and Van Gool *et al.* (2011); and books by Mays (2004), Anderson (2007), Thomas (2009a), van Gestel & Baesens (2009).

3.1 Background

In retail banking, prior to the use of automated credit scoring systems, the credit risk of an applicant was evaluated in a subjective manner based upon underwriters' experiences. Typically, information on the customer was obtained through personal

relationships between the customer and staff at the lender, which curtailed the movement of customers between lenders (Anderson, 2007). Lending was often a judgemental process where an underwriter (typically the bank manager) assessed applications based on criteria known as the 5Cs:

- i. Character - is the applicant or any of their family known to the organisation?;
- ii. Capital - how much of a deposit is the applicant offering and what is the loan amount being requested?;
- iii. Collateral - what security is the applicant offering?;
- iv. Capacity - what is the repaying ability of the applicant?;
- v. Condition - what are the general conditions of the economy at present?

Obviously, such a process had a number of shortcomings, particularly with respect to the consistency and reliability - to put it in a word, quality - of credit granting decisions. Hand (2001) has listed the following key shortcomings: (i) such decisions were undoubtedly affected by the day-to-day changes in the bank manager's mood; (ii) decisions were not always replicable as different managers did not always make the same decisions; (iii) there was no universal formalisation of the decision making process, often making it difficult to teach; and (iv) the human-based judgement approach could only handle a limited number of applications, resulting in lost revenue.

Lending to consumers, however, increased dramatically in the second half of the twentieth century (Thomas *et al.*, 2002). Figure 3.1 illustrates the change in the amount of consumer credit owed to financial institutions in the United States of

America (USA) between 1960 and the first six months of 2012. With the introduction

and subsequent popularity of credit cards (first issued in the USA in 1958 and then in

the United Kingdom in 1966), consumer demand necessitated the development and

growth of objective methods capable of automating the lending decision (Thomas,

2009b). During this period, beginning with the USA, demand for mortgage products

also increased. This demand originated from homeownership policies promoted by

successive American governments after the Great Depression of the 1930s, e.g. the

formation of the Federal National Mortgage Association (or Fannie Mae) as part of

President F.D. Roosevelt's “*New Deal*” (see Romasco, 1983) policies. Motivations

arising from a fear of both communism and labour unrest, ensured that “*stable*

housing was intrinsically linked to the maintenance of a loyal citizenry” [(Wright,

1983) in (Shlay, 2006)]. Indeed, by the 1950s homeownership was identified with the

American Dream to the extent that it is now symbolically equivalent to citizenship

(Shlay, 2006).

Along with the increase in public demand for credit products, regulatory changes

also helped to advance the credit scoring cause. In the interests of fairness and

equality, the United States Congress demanded that the decision-making in credit

granting be made transparent. The Equal Credit Opportunity Act (ECOA), first

enacted in 1974 in the US, and subsequent amendments helped to strengthen the

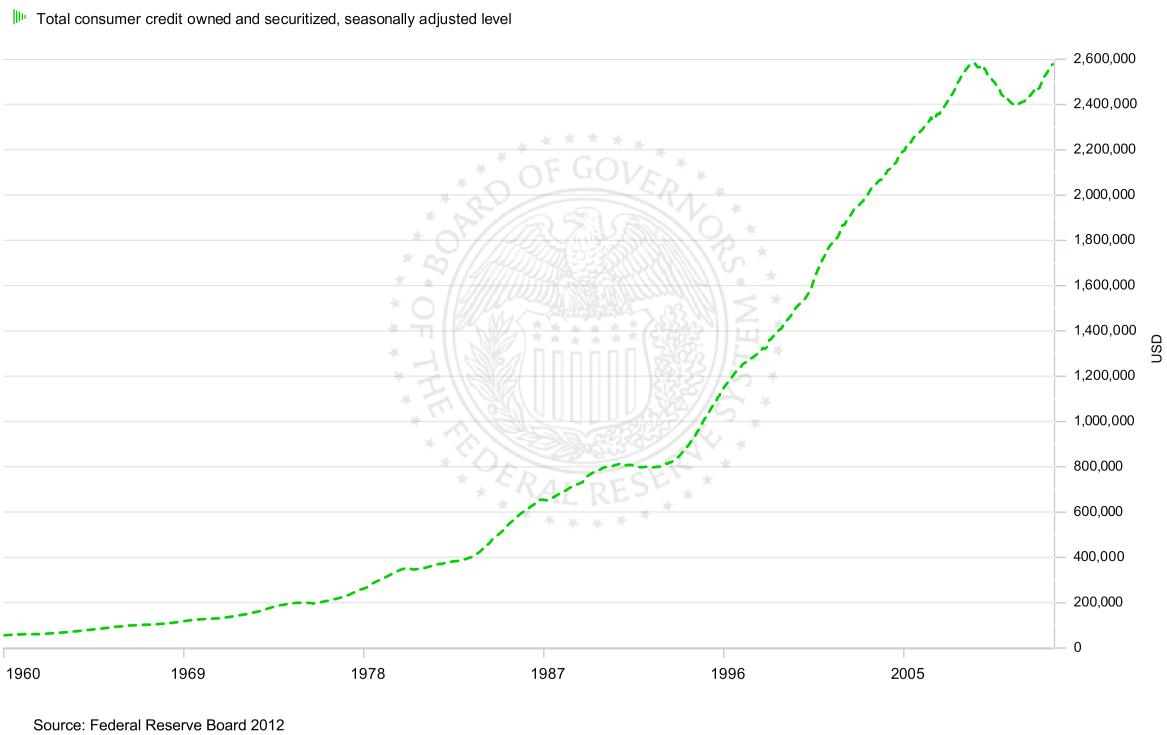
requirement for credit scoring. The ECOA prohibits a creditor from discriminating

against an applicant on the basis of certain prohibited individual details (e.g. race,

colour, or religion).

Statistically developed credit scoring systems were proposed as a means to al-

low creditors to adhere to their regulatory requirements. The US Federal Reserve's



Source: Federal Reserve Board 2012

Figure 3.1: Total consumer credit owned and securitised (seasonally adjusted).
Source: Federal Reserve Board

Regulation B¹ (Section 202.2), which implements ECOA, stipulates that such credit scoring systems must be, amongst other things: (i) “*based on data that are derived from an empirical comparison of sample groups or the population of creditworthy and noncreditworthy applicants who applied for credit within a reasonable preceding period of time;*” and (ii) “*Developed and validated using accepted statistical principles and methodology*”. Although creditors forgo some discretion in their lending, credit scoring systems offer lenders a transparent solution that satisfies the requirements of the ECOA, as they provide a clear explanation to credit applicants when a loan is denied.

Computers provided the necessary means to implement such automated procedures (Hand, 2001). Compared with judgemental schemes, this then resulted in

¹<http://www.federalreserve.gov/bankinforeg/reglisting.htm>

retail banks reporting substantial reductions in (i) the cost of credit evaluations, and (ii) loan losses caused by customer defaults [(Greenspan, 2002) in (Mays, 2004, pp.4)]. In the 1980s with improvements in computational power (e.g. cost, speed, and storage capacity) retail banks began to use statistical methods to monitor, measure, detect, predict, and understand many aspects of customer behaviour (Hand, 2001). Gradually this led to the development of techniques estimating, amongst other things (Hand, 2001; Thomas, 2009a): (i) the risk of default - measuring the risk of a customer defaulting on a particular product (product default scoring) or for any product (customer default scoring) (Hand & Henley, 1997); (ii) fraud detection - techniques that can detect fraud as soon as possible (Phua *et al.*, 2010); (iii) response to advertisement campaigns - will the customer respond to a direct mailing of a new product? (Lee & Cho, 2007); (iv) customer retention - will the customer keep using the product after the expiry of the initial trial period? (Zhao *et al.*, 2005); (v) attrition - will the customer change to another lender? (Thomas, 2001); (vi) product usage - will the customer use a certain product, and if so, to what intensity? (Haenlein *et al.*, 2007); and (vii) profit scoring - techniques to measure the profitability of a customer on a single product (product profit scoring), and over all products (customer profit scoring). Product default scoring is regarded as the original application of credit scoring (Thomas, 2009a). Credit scoring, as previously mentioned in Section 1.2, remains one of the most popular application fields for both data mining and operational research techniques (Baesens *et al.*, 2009).

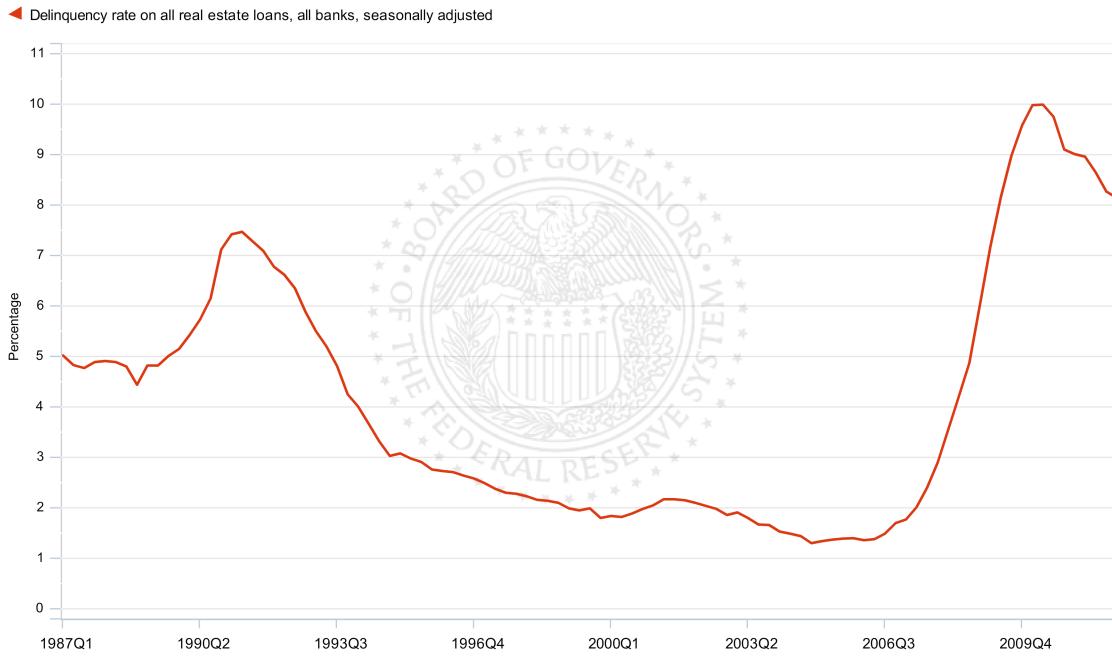
Typically, a credit scoring system is implemented using a credit scorecard. The scorecard assigns points, mechanically, to key customer characteristics and aspects of the transaction in order to derive a numeric value which represents the risk that

a customer, relative to other customers, will default on their financial obligation.

Credit scoring systems are not perfect and can only estimate credit risk based on past, not future, performance. Every year, due to the failure of credit scoring systems to identify individuals who subsequently default on their loan obligation, a significant portion of customer debt goes unpaid (Finlay, 2011). Frequently, the cause of this can be attributed to unforeseen circumstances such as: (i) fraud; (ii) divorce; (iii) financial naivety - lack of financial acumen; and (iv) debt through the loss of income. The delinquency rate (i.e. customers not current with their loan repayments) for residential loans in the United States is displayed in Figure 3.2. As a consequence, there is considerable interest in improving credit scoring systems to discern between profitable and unprofitable customers on the basis of their future repayment behaviour (Finlay, 2011). Amongst practitioners and researchers alike, it is widely accepted that even a small improvement in the assessment of customers' credit risk can result in significant financial savings (Hand & Henley, 1997).

3.1.1 The Basel II Capital Accord

Along with the financial savings afforded by credit scoring there are also regulatory issues to adhere to. In some countries, a central bank is responsible for bank supervision, while other countries have separate —and sometimes multiple —regulatory bodies for bank supervision (Mosley & Singer, 2009). The Bank for International Settlements is an international organisation tasked with promoting international monetary and financial co-operation between central banks. The Basel Committee on Banking Supervision (BCBS) is a subcommittee of the BIS charged with the responsibility for developing guidelines and recommendations on banking regulations



Source: Federal Reserve Board 2012

Figure 3.2: Delinquency rate on all real estate loans, all banks, seasonally adjusted.
Source: Federal Reserve Board

applicable to all member states.

Through the evolution of the Basel Capital Accord (BCBS, 1998; BCBS, 2005a; BCBS, 2010), the BCBS specify an international standard for banks to employ when calculating the necessary amount of capital required to offset potential losses arising from financial and operational risks (i.e. the amount of cash and liquid assets banks must set aside to cover unexpected losses). The first set of proposals, Basel I, developed a set of uniform standards on the level of regulatory capital, and focused principally on credit risk (Wims *et al.*, 2011). Basel I was first published in 1988 and implemented in twelve countries by 1992. It eventually gained universal acceptance as compliant banks received an improved credit rating and lower funding costs (Anderson, 2007). The Basel I regulatory capital calculation uses a straightforward set of rules that assign risk weights to a given asset or loan class. Basel I uses four

broad asset types (sovereign, bank, corporate, and individual) which have different risk weights attached to them. Four basic risk weights were defined, along with an additional category whose weighting is at the discretion of the national regulator, including: (i) 0%, e.g. sovereign debt; (ii) 20%, e.g. debt from other banks or public sector institutions; (iii) 50%, e.g. residential property loans; (iv) 100%, e.g. loans to private sector companies; and (v) 0%, 10%, 20%, or 50% at regulator's discretion (BCBS, 1998). After applying the risk weights to each asset class, the total lending of the asset classes is calculated to provide the sum of *risk weighted assets* (RWA). The required capital ratio is set at a minimum of 8% of the RWA.

A common criticism of Basel I is that it lacked risk sensitivities, affording banks too much flexibility in controlling their RWA via *regulatory arbitrage*. Holman (2010) defines regulatory arbitrage as an activity where “*a bank exploits the difference between its actual risk level and that implied by its regulatory position*”. This practise was performed using complex and opaque financial innovations [e.g. securitisation and credit derivatives (see Kolb & Overdahl, 2009)] which allowed banks to reduce their minimum capital requirements without actually reducing the risk involved. In response to this and other criticisms of Basel I (see Balin, 2008), the Basel II standard then began to evolve from 1999 until the publication of its framework in mid-2004.

Basel II is based upon three mutually reinforcing pillars:

- i. Minimum capital requirements
- ii. Supervisory review process
- iii. Market discipline

The first pillar, minimum capital requirements, describes the methodologies used for calculating and reporting the minimum regulatory capital requirements for credit risk, market risk, and operational risk. The second pillar, supervisory review process, provides guidelines for the supervisory review of the capital adequacy and internal risk assessment processes stipulated in Pillar 1 (van Gestel & Baesens, 2009). The supervisory review process pillar addresses the development and improvement of risk management techniques used to monitor and manage banks' risks. The third pillar, market discipline, attempts to harnesses market discipline to motivate prudent self-regulation by enhancing the degree of transparency in banks' public reporting. As Pillar 1 addresses credit risk, it will be described in further detail in the remainder of this section.

In Pillar 1, the calculation of the minimum capital requirements for credit risk can be performed using methodologies from a continuum of increasing sophistication and risk sensitivity:

- Standardised approach
- Internal Ratings-Based (IRB) approach
 - Foundation approach
 - Advanced approach

Under the standardised approach banks use ratings provided by external credit ratings agencies to quantify the capital requirements for credit risk. Similar to the Basel I framework, the standardised approach uses a risk weighting approach. For increased risk sensitivity, a more detailed classification of the asset classes is defined.

Through the incentive of lower capital reserve holdings, the IRB approaches encourage banks to develop their own internal risk ratings which are capable of measuring banks' actual credit risk. In the IRB approach, both the foundation and advanced approaches are based on four key components:

- Probability of default (PD) is the likelihood that a default event will occur. PD is used as a measure of the borrower's ability and willingness to repay a loan.
- Loss given default (LGD) is defined as the expected economic loss incurred in the case of borrower default. LGD is typically expressed as a percentage of exposure outstanding at the time of default. In the case of no loss, the LGD is equal to zero. Should the bank lose the full exposure amount, the LGD is equal to 100%. A negative LGD would indicate a profit (e.g. due to paid penalty fees and interests on arrears). LGD gives rise to the term *recovery rate*, which can be expressed as $1-LGD$.
- Exposure at default (EAD) is an estimate of the amount owed at default, e.g. the full loan amount plus accrued interest. For certain products such as term loans (or balloon loans) the amount is known before hand. For other products such as revolving credit (e.g. credit cards) the amount varies with the behaviour of the borrower.
- Effective maturity (M) is the length of time before the loan is paid off in full.

As a general rule, under the foundation approach banks provide their own estimates of PD for each asset class, but use estimates provided by regulators for the

other relevant risk components. For the advanced approach, banks must calculate the effective maturity and provide their own estimates of PD, LGD and EAD. However, it should be stipulated that for retail loans there is no distinction between the foundation and advanced approach, and banks must provide their own estimates of PD, LGD, EAD.

At this point it is worthwhile to distinguish between the PD of an individual loan and the PD of a loan portfolio. The PD of an individual loan can be estimated using a classification model, such as any of the ones described in Section 2.2. For example, logistic regression uses the log odds score as a forecast of the PD of individual borrowers, i.e. $\ln\left(\frac{p_g}{1 - p_g}\right)$. The observed PD is then used to assign the customer to a particular rating class. By grouping individual loans, whose PDs are similar, into rating classes an accurate and consistent estimate of the PD of a loan portfolio can be determined.

A loan portfolio consists of individual loans which are grouped together into homogeneous pools. Typically, the segmentation of a retail loan portfolio is performed by, amongst other things: product, acquisition channel, credit score, geographic location, or loan-to-value (Breeden *et al.*, 2008). Lenders may segment a portfolio further by PD bands and, occasionally, LGD bands (i.e. rating class) (Thomas, 2009a). The lender estimates the PD of each rating class, which is the expected number of defaults divided by the number of customers.

Through the use of loan portfolios lenders can utilise the process of securitisation, whereby illiquid assets such as mortgages are transformed into marketable securities. The securities are sold to a third party special purchase vehicle (SPV) who then issue bonds where the loan repayments are used to cover the repayment of the coupons

and principal of the bonds. Through securitisation, lenders can reduce the size of their balance sheet, resulting in lower capital requirements.

The above risk components are used to estimate the expected loss (EL) for each loan portfolio. The EL is a measurement of loss that is anticipated within a one-year period, and is defined as:

$$EL = PD * LGD * EAD \quad (3.1)$$

For example, for a given portfolio, if $PD = 2.5\%$, $LGD = 33\%$, $EAD = €3,000,000$, then $EL = €24,750$. Expected loss can also be measured as a percentage of EAD:

$$EL\% = PD * LGD \quad (3.2)$$

From the previous example, the expected loss as a percentage of EAD would be $0.825\% (2.5\% * 33\%)$.

After deficiencies in Basel II were exposed by the 2008 financial crisis [e.g. insufficient capital requirements, the excessive use of ratings agencies (Wahlström, 2009)], a further revision (Basel III) of the framework was initiated and implementation of the guidelines and recommendations is expected to begin in early 2013. Basel III extends the existing work in Basel II by strengthening capital requirements and introducing requirements on bank liquidity and leverage. As a result, financial institutions must maintain higher capital buffers in order that they are less crisis prone and in need of government bailouts.

To summarise, under the Basel II Capital Accord (BCBS, 2005a), using the internal ratings-based approach, banks can calculate their capital requirements by using their internal data to construct credit risk models. As a consequence of this

approach greater emphasis is placed on an accurate estimation of customers' *probability of default* (PD) rather than the ability to correctly rank customers based on their default risk (Malik & Thomas, 2009). The PD is the "*central measurable concept on which the IRB approach is built*" (BCBS, 2001). PD also has to be predicted not just at an individual level but also for segments of the loan portfolio. A loan portfolio consists of loans segmented into rating bands and the PD is estimated for the customers in each band. Modelling the PD at the loan level is essentially a discrimination problem (good or bad), consequently one may resort to the numerous classification techniques that have been suggested in the literature (e.g. Section 2.2). Many of these classification models are derived from statistical methods, non-parametric methods, and artificial intelligence approaches. By estimating the PD at the account level, and subsequently at the portfolio level, lenders can estimate the loss (or the credit risk) associated with a particular loan portfolio.

The purpose of this section has been to introduce retail credit risk by describing the key drivers behind its establishment and tremendous growth over the second half of the twentieth century. The next section discusses the various stages in developing credit scorecards, which are used to assess the creditworthiness of a customer.

3.2 Credit Scorecards

In its simplest form, a credit scorecard consists of a group of features statistically determined to be predictive in establishing the creditworthiness of a customer (Siddiqi, 2005). The purpose of a credit scorecard is to allow banks to use a structured, transparent, and easy to interpret format with which to assess customers' credit-

worthiness. An example of an application scorecard is displayed in Table 3.1. The table is comprised of *features* and their *attributes*. A feature describes a particular characteristic of the borrower or loan and can be selected from any of the sources of data available to the lender (Siddiqi, 2005). Features which are considered, by statistical means or otherwise, to be predictive of customers' good/bad status are included in the scorecard. Typically, a feature is specified by a group of one or more attributes. An attribute is a member of a set of mutually exclusive values or a range of non-overlapping numbers that the feature can take on. For each attribute, the scorecard assigns a number of points which contribute to an overall credit score. The points assigned to an attribute is based on the analysis of historical data, which involves various factors such as the predictive strength of the feature, the correlation between features, as well as operational considerations (Siddiqi, 2005). The higher the score, the lower the risk of defaulting on a financial obligation.

Table 3.1 includes an example showing how a credit score for a loan applicant, X, would be calculated. The applicant who is 34 years of age, an existing customer with the bank, with a credit card limit of €3,500, 4 years in their current job and is not self-employed, earns a gross monthly income of €3,750, and lives in rented accommodation. Based on this data, the applicant is assigned a credit score of 355 points from a maximum of 443. Lenders look at the score for each applicant and make a decision, based on a cut-off score, as to whether or not to approve the loan. Selecting an appropriate cut-off score is a strategic decision for management involving a trade-off between: (i) expected risk (i.e. predicted bad rate) and return (profit); and (ii) profit and volume (or market share) (Thomas, 2009a).

Although there are a variety of ways a scorecard can be developed, the standard

Table 3.1: Application scorecard with a credit score for applicant X

<i>Feature</i>	<i>Attribute</i>	<i>Points</i>	<i>Attribute value for applicant X</i>	<i>Points for applicant X</i>
Age	< 25	69		
	25 - 29	77		
	30 - 34	84	34	84
	35 - 41	93		
	42 - 50	104		
	50+	110		
Bank Customer	Yes	29	Yes	29
	No	20		
Credit limit on credit card	Blank	60		
	< 2,000	55		
	2,000 - 3,750	59	3,500	59
	3,751 - 6,000	64		
	6,001 - 10,000	71		
	> 10,000	74		
Years at current job	< 1	20		
	1 - 3	24		
	4 - 6	29	4	29
	7+	36		
Accommodation Status	Own	42		
	Rent	28	Rent	28
	Parents	32		
	Other	34		
Self-employed	Yes	25		
	No	41	No	41
Gross Monthly Income	< 2,500	71		
	2,500 - 3,150	79		
	3,151 - 3,850	85	3,750	85
	3,851 - 4,350	92		
	4,351 - 5,100	103		
	> 5100	111		
Score				355

scorecard development process consists of a number of generic stages identified in Section 1.2 as: (i) dataset construction; (ii) modelling; and (iii) documentation. A scorecard development process model is displayed in Figure 3.3. The main tasks performed when developing application and behavioural scorecards are highlighted for each of the three stages. During the dataset construction and modelling stages statistical experts often consult with business experts to assess the consistencies and the variances between empirical findings and business knowledge and experience. At the end of each step, *expert committee approval* (ECA) is required before proceeding to the next step. The expert committee consists of scorecard developers and business experts who meet to review the progress of the development project and determine whether any previous work requires refinement.

Scorecard development is an iterative process and the resultant scorecard must satisfy a number of performance criteria, including (van Gestel & Baesens, 2009):

- Stability - The scorecard attributes should be estimated from a sufficiently sized dataset that covers a suitable historical period.
- Discriminative - The scorecard is expected to distinguish between the goods and bads.
- Interpretable - The output of the scorecard should be understandable and explainable to non-experts.
- Not overly complex - The scorecard should not over-rely on any single feature or consist of too many features.
- Conservative - Basel II requires an estimation of, amongst other things, the

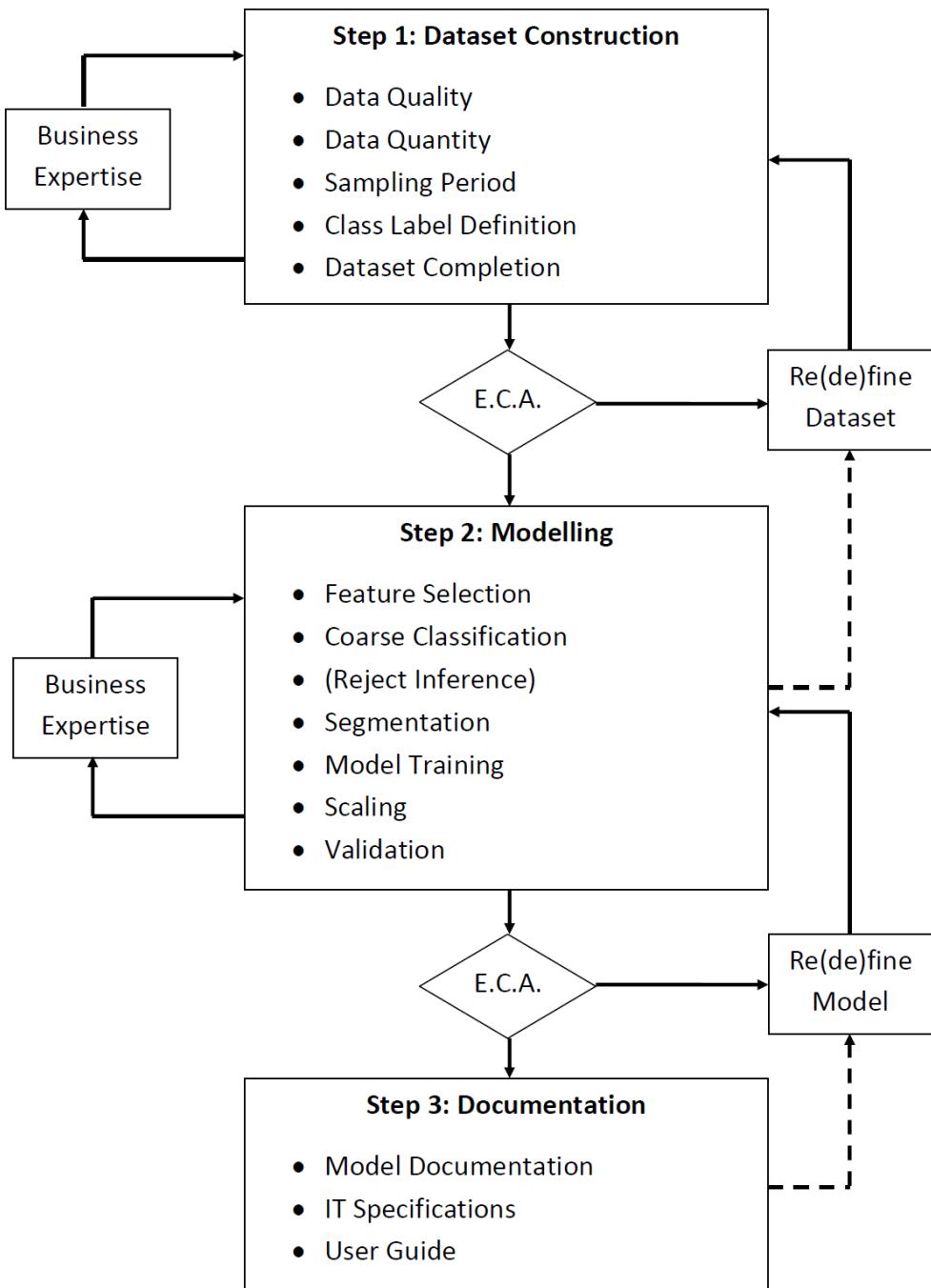


Figure 3.3: A process model for developing a credit scorecard, E.C.A. = expert committee approval. Note: Reject inference is performed during application scoring, but not during behavioural scoring. Based on van Gestel & Baesens (2009).

probability of default of a loan portfolio in order to calculate the amount of capital to set aside to cover losses. Regulatory bodies suggest overestimating this measure to cover downturns in the economy.

- Robust - The scorecard should not exhibit any unnecessary volatility over the economic cycle (i.e. through-the-cycle).

There is no definitive quantitative framework available to ensure that the criterion are satisfied. The decision making process of each financial organisation varies on account of a range of factors such as the available resources and corporate/regional culture. Although regulatory measures such as the Basel Accords provide guidelines, ultimately, a fusion of statistical, legal, information technology, customer, business, and planning expertise is required.

As highlighted previously in Section 1.2, scorecard development relies on successful completion of each of the three stages outlined in Figure 3.3. Of particular interest to this research are the dataset construction, and modelling stages. The documentation stage will not be further investigated in this work. This is not to dismiss it as irrelevant as it is important to record the details of information used, assumptions made, and specifications used for implementation. The rest of this section examines the different tasks performed during the dataset construction and modelling stages.

3.2.1 Dataset Construction

Practitioners often cite the steps performed during this stage as the most time-consuming activities performed during the construction of credit scorecards¹. The main steps performed when creating a dataset with which to construct a scorecard are described hereafter.

3.2.1.1 Data Quality

How successful a scorecard is at discriminating between good and bad applicants depends largely on the data used during the scorecard development stage.

The first issue that needs to be addressed is the quality of data available. In the literature data quality is defined by many characteristics, including accuracy, completeness, and consistency (see Baesens *et al.*, 2009; Lindsay *et al.*, 2010). *Data accuracy* relates to the degree of precision of measurements of a feature to its true value (Baesens *et al.*, 2009). Common examples listed among the typical causes of poor data accuracy are user input errors and errors in software. *Data completeness* refers to the extent to which values are missing in the data (Parker *et al.*, 2006). *Data consistency* relates to situations in which multiple data sources are used and due to a lack of standardisation, two or more data items may conflict with each other. Although there are various methods used to handle missing or incomplete data (see Florez-Lopez, 2009), the simplest approach is to remove the affected entries from the dataset.

¹For example see http://www.kdnuggets.com/polls/2003/data_preparation.htm

3.2.1.2 Data Quantity

To ensure the construction of a high quality and robust scorecard a sufficient quantity of customer data is required. During this step data sources are identified and guidelines are established as to how the data may be procured. The amount of data necessary depends on the objective of the scorecard and the properties of the data with respect to the scorecard objective. Traditionally, in application scoring, industry experts recommended using 1,500 instances of each class [(see Anderson, 2007; Lewis, 1992; Siddiqi, 2005)]. Where reject inference (see Section 3.2.2.3) is performed, an additional 1,500 rejected applicants were required.

Anderson (2007) attributes these numbers to the 1960s, an era when computational power was limited and the collection of data was more costly. Today, these quantities are still largely used, and for many practitioners the validity of the recommendations is based on the understanding that the composition of credit scoring datasets is homogeneous across lenders and regions (Crone & Finlay, 2012). For example, consider the different sources of credit scoring data:

- Internal - This type of data details past customer dealings and other account behaviour. Account balance, years as customer, and existing loans with the bank are examples of customer account information that is stored internally by the bank.

- External - This type of data is obtained from application forms and financial statements. Examples of such information include: number of dependants, number of years at current address, and income.

- Bureau - This relates to data held by credit bureau and court records. Credit bureaus are institutions that collect data on the performance of loans granted by different lenders. Some credit bureaus also detail the number of loan applications that were submitted during the last 12 months.

To capture this data the majority of lenders ask similar questions and use standardised industry data sources supplied by credit bureaus, resulting in features and properties that are broadly similar (Crone & Finlay, 2012). Another reason for such recommended sample sizes can be attributed to the fact that they are sufficiently large to exhibit the same properties as the population of interest. Finally, the dataset should contain enough instances to restrict the occurrence of correlated variables which can result in over-fitting of the scorecard model.

3.2.1.3 Sampling Period

As mentioned previously, credit scorecards are built using historical data. Although past performance does not guarantee future performance, in credit scoring, history is a reliable indicator. To generate application scoring training datasets a snapshot of each customer is taken at two different points in time (Martens *et al.*, 2010). The first occurs at the beginning of the loan when the customers' characteristics are recorded. The second occurs sometime later at the *default observation point*, at which point the customer is classified as either good or bad. The period of time between the two snapshots is commonly referred to as the *outcome window*. During this step the size of the outcome window is specified, based on the business objectives of the scorecard. For example, a short outcome window (e.g. 6 months) may be used when the goal is detect defaulters as soon as they have fallen into arrears

without accounting for the possibility that the borrowers may recover. In order to limit the chances of misclassifying a customer and to avoid understating the default rate lenders must decide on an appropriate length of the outcome window. If the outcome window is too long there is a possibility that differences may arise between the sample used during scorecard development and as-yet unseen future samples. Such differences may arise from changes in macro-economic conditions, company strategy, and personal circumstances (Hoadley, 2001). If the outcome window is too short valuable information may be lost, for example, certain default events may not have occurred. Typically, for mortgages, the outcome window is identified by plotting the monthly cumulative default rate. The cumulative default rate is calculated as the total number of borrowers divided by the total number of defaults. A plateau in the monthly cumulative default rate indicates the maturity of the data sample. One would expect a plateau in the default rate to occur after three to five years (Siddiqi, 2005). However it is not uncommon to select an outcome window prior to this period, provided that the discriminatory power and intuitiveness of the credit scorecard is not affected.

3.2.1.4 Class Label Definition

How a loan account is defined as bad depends on the objectives of the scoring system and the financial institution's view of success or failure (McNab & Wynn, 2000). Typically, in credit scoring, a loan account is labelled as bad when a default has occurred. Using the Basel II definition (paragraph 452), a default is considered to have taken place when either or both of the following criteria are fulfilled:

- the borrower is past due more than 90 days on any material credit obligation

to the lender.

- the lender considers that the borrower is unlikely to repay its credit obligations to the lender in full, without recourse by the lender to actions such as realising security (if held), e.g. home repossession in the event of loan default

According to Anderson (2007), financial institutions can choose between: (i) a *current status* label definition approach that classifies a customer as either good or bad based on their account status at the end of the outcome window; and (ii) a *worst status* label definition approach which classifies a customer as either good or bad based on their account status during the outcome window. Commonly, as per Basel II (BCBS, 2005a), a customer's 90-days *worst status* covering a one-year period is considered an appropriate definition for bad accounts. The *current status* label definition, however, is often used when managing early-stage delinquencies (Anderson, 2007).

3.2.1.5 Dataset Completion

The final step of the dataset construction stage involves splitting the data into two portions: the training sample and the testing sample. The training sample is used to build the scorecard and the testing sample estimates how well the scorecard performs. There are various ways to split the training and testing datasets. Normally, where there is sufficient data available, scorecard builders opt for the hold-out approach (see Section 2.5.2) with a 70:30 split between the size of the training and testing samples (Siddiqi, 2005). In the hold-out approach, a portion of the training sample, called the validation sample, is set aside for tuning the parameters of the

underlying scorecard model.

Where there is insufficient data available, standard statistical approaches such as cross-validation (see Bishop, 2006) and bootstrapping (see Japkowicz & Shah, 2011) can be used to estimate model parameters without losing too much information (Thomas, 2009a).

3.2.2 Modelling

After the training and testing datasets have been generated, the scorecard can be developed in the modelling stage. There are a number of different steps to perform during the modelling stage, each of which is described hereafter. Much of the discussion that follows is based on the assumption (highlighted in Section 2.2.2) that logistic regression is perhaps the most commonly used algorithm within the consumer credit scoring industry.

3.2.2.1 Feature Selection

Feature selection is the process of choosing a subset of the full set of features available for use in a scorecard by eliminating features that are either redundant or possess little predictive information. In the literature, the topic of feature selection has been discussed extensively (see Guyon & Elisseeff, 2003), but briefly, feature selection techniques are commonly categorised into one of three groups: (i) filter techniques; (ii) wrapper techniques; and (iii) embedded techniques. Filter techniques assess the relevance of features using only the intrinsic properties of the data and are independent of the classification algorithm. With wrapper techniques various subsets of features are generated and evaluated using a specific classification algorithm. Wrap-

per techniques combine a specific classification algorithm with a strategy to search the space of all feature subsets. In the third category of feature selection techniques, termed embedded techniques, the feature selection strategy is built into the classification algorithm. In this thesis, we restrict the rest of our discussion to commonly used feature selection techniques employed in credit scoring.

In credit scoring there is usually a large set of candidate features emanating from the variety of sources used to record customer and macroeconomic environment characteristics (e.g. see Section 3.2.1.2). A robust scorecard typically uses between 10 and 20 features (Thomas, 2009a), although Mays (2004) recommends between 8 and 15. There are a number of valid reasons for performing feature selection during scorecard construction. Firstly, from a practical perspective, in order to reduce costs it is important to remove as much irrelevant and redundant information as possible. Otherwise, staff are paid to analyse and understand additional features that are effectively redundant when assessing customers' creditworthiness. Secondly, identifying predictive features assists in providing clearer insight into, and a better understanding of, the scorecard. Finally, by applying the principle of Occam's razor, a simple scorecard with optimal predictive accuracy is preferable to a more complex scorecard that includes many unnecessary features. The *curse of dimensionality* is the term used when too many irrelevant and redundant features and not enough instances describe the target population (see Loughrey & Cunningham, 2005). This can result in over-fitting, whereby the induced model accurately classifies the instances in the training sample, including the noisy ones, but performs poorly when applied to a previously unseen sample.

Feature selection is affected by the following factors (Mays, 2004; Siddiqi, 2005):

(i) cost; (ii) legal; (iii) business logic; and (iv) statistical analysis. The cost factors include the computational and financial costs involved in acquiring the input. The legal factors relate to the use of features that pose a legal, regulatory, or ethical concern. Credit scorecard builders must ensure that the features used are compliant with any of these concerns. The situation will vary from country to country. For example, in the United Kingdom (UK), Section 29 of The Sex Discrimination Act 1975 prohibits the use of features, such as gender, that are discriminatory in the granting of credit against members of a protected class of customers. By contrast, Mwangi & Sichei (2011) illustrate that in Kenya, gender is often used in credit scoring as it has a direct relationship to credit access.

Using business logic, scorecard builders can justify the inclusion and removal of certain features. With the expert knowledge acquired from previous scorecard development projects, practitioners select certain features for inclusion because of their expected predictive power. For example, a particular feature may reveal certain idiosyncrasies of a sub-population. Business logic may also be used to judge the reliability of features. For instance, certain commission-based sales agents may manipulate unconfirmed data to increase an applicant's chances of being granted credit. Business logic can flag feature values considered unusual for an instance belonging to a certain sub-population. Business logic can also be used to determine the future stability and availability of features as it is important that features used in the initial dataset should also be available for future samples. Finally, to avoid the overuse of ratios business logic should be used to justify their inclusion. Ratios are constructed by combining existing features, a consequence of which may be an increased incidence of correlated features (Anderson, 2007). When the intercorrelations among

features are very high, this is likely to cause multicollinearity problems, which can result in poor scorecard performance on previously unseen data (Diamantopoulos & Siguaw, 2006).

Statistical analysis is the final factor used in variable selection. Statistical analysis techniques are used to identify highly correlated features which must be removed in order to determine the true contribution of each feature to the class label. Three commonly used techniques include [(Morrison, 2004) in (Leung Kan Hing, 2008)]:

- Correlation-based feature selection (CFS)
- Stepwise procedures
- Factor analysis

Correlation-based feature selection: CFS methods are an example of filter feature selection techniques. Two straightforward CFS methods often used in conjunction with each other are bivariate and pairwise correlation. Bivariate correlation measures the relationship between each feature and the class label. Pairwise correlation measures the relationship between each of the features. A correlation matrix, similar to Table 3.2, is constructed containing both the bivariate and pairwise correlations. A pairwise correlation threshold (e.g. 0.70) is used to identify candidate features for removal. For features whose pairwise correlation is above the correlation threshold (e.g. Expenses and Income in Table 3.2), the feature with the lowest bivariate correlation value is removed (i.e. Expenses, 0.12). Although this technique is easy to implement, computationally fast, and scales easily to high dimensional data; it does not perform any tests of statistical significance and only one pair of elements is examined at a time (Leung Kan Hing, 2008). Refer to Atiya

(2001) for an example of a correlation matrix applied to a bankruptcy prediction dataset.

Table 3.2: Correlation matrix for bivariate and pairwise correlation. The diagonal of the matrix has values of 1.00 because a variable always has a perfect correlation with itself.

	<i>Expenses</i>	<i>Income</i>	<i>Loan Value</i>	<i>Class Label</i>
<i>Expenses</i>	1	0.71	0.27	0.12
<i>Income</i>	-	1	-0.34	0.56
<i>Loan Value</i>	-	-	1	0.88
<i>Class Label</i>	-	-	-	1

Stepwise procedures: The second technique, stepwise procedures, is a wrapper feature selection technique, and consists of iteratively performing linear or logistic regression on the class label using a subset of the features (Thomas, 2009a). Regression techniques address a shortcoming with correlation methods by evaluating the correlation between features collectively, rather than one pair at a time. The objective of the stepwise procedure is to identify the most parsimonious set of features that adequately describe the class label (Hosmer & Lemeshow, 2000). Features are added and (or) removed from the regression model using techniques such as forward selection (see Hocking, 1976), backward elimination (see Myers, 1990), and forward-backward selection (see Pearce & Ferrier, 2000).

In forward selection, the model initially contains no features, and features are added incrementally until a final model is obtained. In backward elimination, all features are included in the initial model, the features are then removed incrementally until a final model is obtained. Forward-backward selection is a combination of the previous two techniques, in which each forward step is followed, though not necessarily, by a backward step to remove the least predictive feature(s) in the model.

Factor analysis: The third statistical analysis technique, factor analysis, is used to transform a large set of correlated features into a smaller set of latent underlying factors. The goal of factor analysis is to obtain parsimony by using the fewest possible uncorrelated features to explain the maximum amount of common variance in a correlation matrix (Tinsley & Tinsley, 1987). In factor analysis, features are divided into common factors and unique factors (Diamond & Simon, 1990). A common factor refers to a latent feature that accounts for variance shared by multiple observed features. A unique factor refers to a latent feature that accounts for variance of one of the observed features not shared with any other observed feature. Unique factors are not related to common factors or to other unique factors. Principal axis factoring and maximum likelihood are two common methods used to perform factor analysis.

The obvious advantage of factor analysis is the reduction of the number of features by combining multiple features into a single factor. Furthermore, the resulting factors are uncorrelated features which may account for much of the variability in the original data (Park *et al.*, 2005). However, in practise, the resulting factors may not necessarily be completely uncorrelated as this is dependant on the factor rotation and score extraction methods used. Lastly, the factors maybe be difficult to interpret in a meaningful way, or conflicting interpretations may arise (Ozkaya & Siyabi, 2008; Yanovskiy *et al.*, 2007).

Other popular feature selection approaches suggested in the credit scoring literature include: (i) variable clustering (see Leung Kan Hing, 2008); (ii) partial least squares (see Yang *et al.*, 2011); (iii) a form of factor analysis called principle component analysis (see Canbas *et al.*, 2005; Min & Lee, 2005); (iv) univariate statistical

analysis such as t-tests (see Huang *et al.*, 2004; Shin *et al.*, 2005); (v) variable ranking using chi-square statistic, Spearman rank-order correlation, and information values (see Section 3.2.2.2); and (vi) statistical learning techniques including, amongst others, decision trees (see Ratanamahatana & Gunopulos, 2003), genetic algorithms (Shin & Lee, 2002), and neural networks (see Castellano & Fanelli, 2000).

3.2.2.2 Coarse Classification

After the number of features have been reduced to a manageable level, the next step is to transform the data into a form appropriate for the scorecard modelling process. Data transformations are often employed to simplify the structure of the data in a manner suited to the modelling (Carroll & Ruppert, 1988). The most commonly used approach in credit scoring is coarse classification. For continuous features, coarse classification codes values into a small number of categories. Similarly, the attributes of categorical and ordinal features are often aggregated into a smaller number of categories. This allows each category of each feature to be treated as a dummy feature having its own weight in the industry standard logistic regression model (Hand *et al.*, 2005). In contrast, with a continuous feature, a single regression coefficient is estimated which may not adequately capture the feature's non-linear relationship with the class label.

Coarse classification increases a scorecard's robustness by reducing the possibility of over-fitting and creating categories with sufficient numbers of good and bad observations (Baesens *et al.*, 2009). Another benefit occurs when certain features exhibit a non-monotonic relationship with the likelihood of default. For example, consider the feature "*time living with parents*". In the USA, anecdotal evidence

suggests that borrowers over the age of 30 living with their parents are deemed of having a greater risk of default as it is regarded as the norm to leave home by that time (Siddiqi, 2005, pp.49). However, if the borrower leaves home at too early an age (e.g. 18-19 years of age), it may indicate a lack of savings due to paying rent and other household bills and so increase the risk of default. Coarse classification is used to accommodate such non-linear relationships by creating several separate categories, each of which has its own weight in the standard logistic regression model. Another advantage of using coarse classification is the ability to incorporate missing values by using a separate category. Similarly, the instabilities caused by outliers and extreme values can be addressed, in part, by aggregating such values into a separate category. In the literature coarse classification is also referred to as *binning*, *grouping* and *discretisation*.

The standard approach to coarse classification is to split each feature into approximately three to six categories (Hand *et al.*, 2005). The model may become over-parameterised and unwieldy if more categories are used. Conversely, the model becomes inflexible when fewer are used. When coarse classifying a categorical feature, attributes with approximately equal good-to-bad ratios are grouped together into coarse classes (Thomas, 2009a). Typically with ordinal features, adjacent attributes are banded together. For continuous features, an initial division of the values into 10-20 categories is performed using the range between minimum and maximum values (Lin *et al.*, 2011). Similarly to ordinal features, adjacent groups are then banded together to produce a smaller number of coarse categories with similar good-to-bad ratios. Regardless of the data type, the literature recommends that categories are sufficiently large enough to contain at least 5% of the sample

population (Thomas, 2009a). Anything smaller may lead to unreliable estimates of scorecard attributes.

Estimating the weight of evidence (WoE) of each category, and fine tuning as required, is the most commonly applied approach to perform coarse classification in credit scoring (Thomas, 2009a). The weight of evidence of category i is defined as:

$$\text{WoE} = \ln \left(\frac{n_g(i)}{n_b(i)} \middle/ \frac{N_g}{N_b} \right) \quad (3.3)$$

where N_g and N_b are the total number of goods and bads in the data sample. The number of goods in category i is $n_g(i)$, and the number of bads is $n_b(i)$. Table 3.3 shows an example of how the WoE of a feature is calculated. To calculate the WoE of a feature, the observations are grouped into equal sized groups (column 1) based on their feature values. The log odds of each group is then calculated (column 7). This is the log of the ratio of the number of goods (column 2) to the number of bads (column 3) for each group. The WoE for each group is then calculated by subtracting the log odds of each group (column 7) from the log odds of the overall population (i.e. 4.34). A negative WoE indicates that the particular group is more likely to default on their loan obligation; a positive WoE indicates the reverse.

The information value (IV) is often used in conjunction with the weight of evidence. The IV indicates the predictive power of a feature and is defined as:

$$\sum_{i=1}^j \left(\frac{n_g(i)}{N_g} - \frac{n_b(i)}{N_b} \right) * \text{WoE}_i \quad (3.4)$$

where j is the number of categories. The result for each attribute is known as the

contribution, which are then summed together to give the IV of a feature. The IV is also referred to as the Kullback divergence measure, and is used to measure the difference between two distributions (Anderson, 2007). Based on Table 3.3 to calculate the IV of a feature, first calculate the percentage of goods (column 4) and the percentage of bads (column 5) in each group/attribute. Next, calculate the contribution of a group by multiplying the respective WoE (column 8) by the difference between the percentage of goods and the percentage of bads (column 6). Sum the contributions (column 9) together to get the IV of the feature. Generally, characteristics with an IV greater than 0.3 are considered highly predictive (Mays, 2004). An IV greater than 0.5 may be too predictive and should be investigated further to avoid over-fitting (Siddiqi, 2005). A IV of less than 0.1 is considered weak and is a candidate for exclusion from the scorecard (Anderson, 2007).

Table 3.3: Analysis of a grouped feature. G = goods, B = bads.

1 Group	2 $\# G$	3 $\# B$	4 $\% G$	5 $\% B$	6 $\% G - \% B$	7 $\ln\left(\frac{\#G}{\#B}\right)$	8 WoE	9 IV
1	3,041	24	10.05%	6.08%	3.98%	4.84	0.50	0.0200
2	3,047	18	10.07%	4.56%	5.51%	5.13	0.79	0.0437
3	3,042	23	10.05%	5.82%	4.23%	4.88	0.55	0.0231
4	3,036	29	10.03%	7.34%	2.69%	4.65	0.31	0.0084
5	3,029	36	10.01%	9.11%	0.90%	4.43	0.09	0.0008
6	3,016	49	9.97%	12.41%	-2.44%	4.12	-0.22	0.0053
7	3,021	44	9.99%	11.14%	-1.15%	4.23	-0.11	0.0013
8	3,013	52	9.96%	13.16%	-3.21%	4.06	-0.28	0.0089
9	3,017	48	9.97%	12.15%	-2.18%	4.14	-0.20	0.0043
10	2,993	72	9.89%	18.23%	-8.34%	3.73	-0.61	0.0509
Total	30,255	395	100%	100%		4.34		0.1669

Through a process of experimentation and fine-tuning different groupings can be combined to surpass some specified minimum predictive strength as measured using the WoE and IV. In doing so, a number of factors need to be taken into consideration. Firstly, the more the log odds of a group differ from the log odds of all groups,

the greater the absolute value of the group's WoE. To maximise the differentiation between goods and bads the absolute value of the WoE of each category should contain enough observations to indicate predictive strength. In the literature, very few recommendations are provided on the number of observations to use in each group. The resulting IV should indicate whether or not further observations are required. Secondly, and just as importantly, the difference between the WoE of categories should be large enough to ensure an acceptable predictive strength of the actual feature (Siddiqi, 2005). The utilisation of expert knowledge is also required during the process as the defined categories should follow some logical trend or have some logical relationship. Such actions improve the interpretability of a scorecard making it easier for financial institutions to understand and explain their decisions to customers (Lin *et al.*, 2011).

The chi-square statistic (see Thomas *et al.*, 2002, pp.132) and Somer's D concordance statistic (see Thomas *et al.*, 2002, pp.134) can also be used during coarse classification to identify the optimal groups.

3.2.2.3 Reject Inference

In application scoring when constructing a scorecard, the outcome value (i.e. good or bad) is only available for customers who were actually granted credit. For customers who were declined credit - as they were deemed to represent a default risk - one only has their characteristic values but not their outcome data. This is a form of sample bias, often referred to as *reject bias* (see Thomas *et al.*, 2002), where the bank's customer database is not representative of the *through-the-door* applicant population (Chandler & Coffman, 1977). Reject inference (see Hand & Henley, 1993) attempts

to address this bias by estimating how rejected applicants would have performed had they been accepted. By using reject inference techniques practitioners attempt to (Thomas, 2009a): (i) improve the discrimination of the scorecard; and (ii) provide an accurate estimate of scorecard performance on the actual application population to which it will be applied, rather than only the accepted population.

The simplest approach to dealing with reject bias is to obtain the customer features and outcome for the entire applicant population by granting credit to every applicant during some time period. Traditionally, retailers and mail order firms have used this approach (Thomas *et al.*, 2002). However, for many banks this is financially infeasible given the losses that are likely to occur.

A number of different reject inference approaches have been developed to address this bias. A crude way is to simply designate each rejected applicant as bad. Obviously, a drawback with this approach is that it reinforces the bias of previous decisions given that some group of customers could be labelled as bad without the chance of disproving this assumption (Thomas *et al.*, 2002). Two of the most commonly used reject inference approaches are *extrapolation* and *augmentation*. There are several variants to each approach. Extrapolation (see Meester, 2000) is a relatively simple method that estimates a preliminary model using only the accepted applicants. Next, this model is used to extrapolate the probability of default for the rejected applicants which is used to impute a good–bad classification to the rejected applicants based on a cut-off probability. Finally, a new model is estimated using both rejected and accepted applicants. With the augmentation approach (also known as re-weighting), a model is estimated using the accepted applicants but each applicant is weighted by the inverse of the probability of being accepted. To calcu-

late this inverse probability a second model is estimated using both the accepted and rejected applicants and this model predicts which applicants will be accepted (Mays, 2004). Based on this approach, in order to simulate the presence of rejected applicants, a disproportionately higher weight is given to the more marginal customers (Banasik & Crook, 2009).

Alternative reject inference techniques discussed in the literature include: multiple imputation (Fogarty, 2006), mixture methods (Feeiders, 2000), iterative reclassification (Joanes, 1993), bivariate probit with sample selection (Banasik *et al.*, 2003), bound and collapse methods (Sebastiani & Ramoni, 2000), a modified logistic regression method (Chen & Astebro, 2006), a bound and collapse Bayesian technique (Chen & Åstebro, 2011), and using survival analysis to reclassify rejects (Sohn & Shin, 2006).

Amongst the credit scoring community there seems to be little agreement as to the improvements (or indeed lack of improvements) associated with using reject inference techniques. It is also unclear which is the best technique to handle reject inference. In part, this may be attributable to the relative lack of empirical studies on datasets that include results for both accepted and rejected applicants. This makes it difficult to measure the significance of reject bias. Crook & Banasik (2004) report that only when a very large proportion of applicants are rejected, is there scope for, at best, modest improvement in scorecard performance through the use of reject inference. Indeed, the same authors reported that extrapolation appears to be both useless and harmless, and re-weighting appears to perform no better than an unweighted estimation of regression parameters.

3.2.2.4 Segmentation

An early decision in scorecard modelling is whether or not to segment the population and build different scorecards for each segment. Segmentation is performed by dividing the population into several groups and building separate scorecards for each group. In marketing, Wedel & Kamakura (2000) describe how segmentation is regularly used to group customers into homogeneous groups based on their purchasing patterns and demographic information such as, amongst others, income, and age (Hand *et al.*, 2001). In credit scoring, the purpose of segmentation is to improve scorecard discriminability and allow greater lender flexibility with regard to product configurations such as interest rate, repayment structure, and other such requirements. The construction and maintenance of additional scorecards involves additional labour and requires careful consideration to limit the number of segments. The three considerations influencing the decision to segment the data are (Thomas, 2009a): (i) operational; (ii) statistical; and (iii) strategic.

Operational considerations are concerned with the acquisition and availability of data which is responsible for differences between the segments (Anderson, 2007). For example, younger customers may have less historical data than more established customers. Another example is whether or not the applicant has a current account with the bank and a subsequent record of customer behaviour. Operational considerations may include certain biases arising from data which originates from different channels (e.g. internet customers, mortgage brokers) where the level of third party advice to the applicant may vary. Finally, mergers between banks can also result in substantial differences in customer details resulting in separate scorecards for each

customer group.

The statistical considerations concern highly predictive features that interact strongly with one another. Two features interact with one and other when the predictiveness of one feature varies based on the value of the other feature (Anderson, 2007). For example, the risk associated with marital status may vary depending on the number of children (e.g. a single parent is often deemed riskier than a couple with children). To limit the inclusion of too many interacting features, scorecard builders often prefer to construct separate scorecards for each attribute of a highly predictive feature.

The strategic considerations relate to policy decisions the bank may wish to implement. For example, wealthier customers may receive a preferential rate of interest on their loan. Segmenting the scorecard population in this manner makes it easier for the bank to manage its customers by employing strategies best suited to those customers.

To perform segmentation, scorecard builders employ both experience-based and statistical approaches (Siddiqi, 2005). Experienced-based approaches rely on the application of business knowledge and industry practices to identify homogeneous sub-populations with respect to some feature. Statistical approaches use statistical tools and statistical learning techniques to identify suitable segments of the population. For example, cluster analysis techniques such as K -means clustering and self-organising maps are used to establish different groups based on certain customer characteristics. Commonly, after the data has been segmented, logistic regression is then employed in the normal manner to construct a scorecard for each segment. A drawback with this approach is that the identified groups may not differ in risk pro-

files as the customers' class label (good or bad) is not used during the segmentation process. Furthermore, the initial segmenting used at the beginning of the iterative segmentation process is based on random vectors, which may greatly affect the final outcome, resulting in local, rather than global optimum (Sherlock *et al.*, 2000). This can be addressed using tree structured classification [e.g. classification and regression trees (CART) (Breiman *et al.*, 1984)] which use the customers' class label to isolate segments. For example, using 3 real world datasets, Bijak & Thomas (2012) evaluated the use of CART in addition to two other segmentation approaches - Chi-squared Automatic Interaction Detection (CHAID) trees (Kass, 1980), and Logistic Trees with Unbiased Selection (LOTUS) (Chan & Loh, 2004). The authors reported that the suite of segmented scorecards did not perform considerably better than the single-scorecard system.

To justify the extra costs involved in the development, implementation, maintenance, and monitoring of a suite of scorecards, the data needs to be "*sufficiently different*" and large in size (Banasik *et al.*, 1996). Further challenges arise when too few bads occur within each segment for reliable scorecard validation (Mays, 2004). If these criteria are met then the extra costs associated with multiple scorecards should be compensated for by the improvement in performance. However important model performance is, segmentation is sometimes driven by operational and strategical factors similar to those previously described.

3.2.2.5 Model Training

With the creation and processing of the training and testing datasets, the training of the predictive model can begin. It is standard practice in the industry to use logistic

regression at this stage of the scorecard development. As highlighted previously in Section 2.2.2, logistic regression is perhaps the most commonly used algorithm within the consumer credit scoring industry (Hand & Zhou, 2009). Any of the predictive modelling techniques described in Section 2.2 may also be considered. A predictive model is trained using the training dataset and the testing dataset is used to assess the accuracy of the model. Whilst it is important that the model separates the goods and the bads, it is also necessary to consider how well the model fits the data in order to avoid problems such as over-fitting (as described in Section 3.2.2.1). Depending on the performance of the predictive model, the datasets may need to be revised by revisiting steps from the dataset construction stage. Performance is often measured using the Gini coefficient (see Section 2.5.1.2) and the Kolmogorov–Smirnov (KS) statistic (see Hand, 2012).

3.2.2.6 Scaling

The predictive models described in Section 2.2 output a probability that can be translated into either a good or bad class. For example, binary logistic regression uses the probability of class membership to express the log likelihood ratio of good-to-bad (or the log odds) in the form of (repeating Equation 2.6):

$$\ln(\text{odds}) = \ln\left(\frac{p_g}{1 - p_g}\right) = b_0 + b_1x_1 + \dots + b_nx_n \quad (3.5)$$

In credit scoring, it is common practice to prescribe a score to such probabilities, e.g. Fair Isaac Corporation (FICO) credit scores range from 300 to 850¹. The bank

¹<http://www.myfico.com/CreditEducation/articles/>

then decides on a cut-off score so those with scores below the cut-off are classified as undesirable and those with scores above the cut-off are classified as desirable (Thomas *et al.*, 2001a).

Scorecard scaling is used to transform the output of a predictive classifier to a score which represents a particular good-to-bad ratio. Scaling does not affect the predictive strength of the scorecard (Siddiqi, 2005). Instead, scaling is a cosmetic exercise performed, primarily, to improve the ease of understanding and interpretation of a scorecard to non-expert users. A survey conducted by Thomas *et al.* (2001a) identified a number of desirable scorecard properties including: (i) the total score is positive; (ii) the points for each scorecard attribute are positive; (iii) there are reference scores which have specific good:bad odds; (iv) the differences between scores has a constant meaning throughout the scale.

Scaling can be implemented using a variety of approaches (see Thomas *et al.*, 2001a), one such approach, linear scaling, is described using:

$$\text{Score} = \text{Offset} + \text{Factor} * \ln(\text{odds}) \quad (3.6)$$

where $\ln(\text{odds})$ is the log odds score calculated using Equation 3.5. The Factor represents the number of points, y , required for the odds to increase by some specified multiple, m , and is defined as:

$$\text{Factor} = y / \ln(m) \quad (3.7)$$

For example, as it is common for the odds to double every 20 points [as per (Siddiqi,

2005, pp.114) and (Thomas, 2009a, pp.43)], then the Factor is calculated as:

$$20/\ln(2) = 28.85$$

The Offset is the base point, b , at which some specified odds, j , occur and is defined as:

$$\text{Offset} = b - (\text{Factor} * \ln(j)) \quad (3.8)$$

Using the Factor value calculated above, the Offset for odds of 30:1 at 200 points¹ is calculated as:

$$200 - (28.85 * \ln(30)) = 101.88$$

The score corresponding to each set of odds (or attributes) can now be calculated using Equation 3.6 as:

$$101.88 + (28.85 * \ln(\text{odds}))$$

Figure 3.4 displays the scaled score for the above example. At $\ln(30)$, or 3.4, the scaled score is re-calibrated to 200. As the odds double the scaled score increases every 20 points, e.g. $\ln(60)$, or 4.09, the scaled score is 220.

On account of the variety of classification approaches available in the modelling stage along with the flexibility of approaches in the dataset construction stage score-card builders often construct at least two or three different scorecards (Siddiqi, 2005, pp.119). Selecting a final scorecard involves the use of the evaluation measures described in Section 2.5.

¹these are arbitrarily selected values used for guidance in the current example

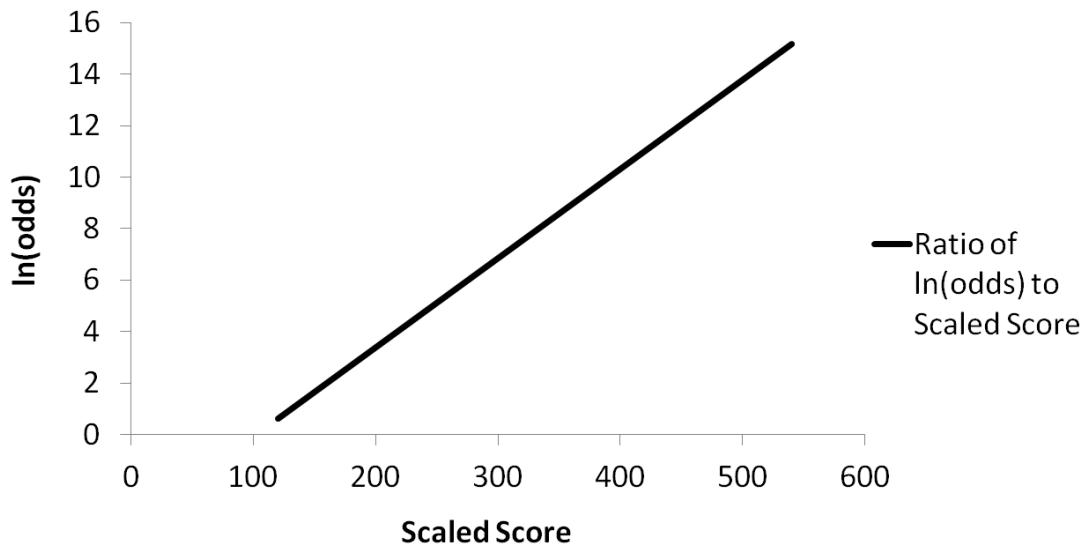


Figure 3.4: Scorecard scaling using linear scaling.

3.2.2.7 Validation

Prior to implementation, the credit scorecard must be evaluated for consistency and accuracy. This process is referred to as *validation* and is usually undertaken by a party independent of the development process and performed using data which was not used in the modelling stage. Validation is an ongoing activity, performed not only after the scorecard has been developed, but also on a periodic basis, especially when any significant structural changes to the scorecard have occurred. A requirement of Basel II is a suitable process to validate the estimates of PDs, LGDs and EADs. The validation of PDs involves two broad dimensions: *discriminatory power* and *calibration* (Stein, 2002).

The discriminatory power refers to degree of separation of the distributions of scores between the goods and bads (Crook *et al.*, 2007). A Confusion matrix and the performance measures derived from it are normally used to measure the power of a model. Other commonly used tools for evaluating credit scoring models include

power curves (e.g. ROC curve, Kolmogorov-Smirnov statistic) and Cumulative Accuracy Profile (CAP) plots (see Sobehart *et al.*, 2000) which graphically illustrate the power of various models on a dataset (Stein, 2002).

Calibration refers to the accuracy of the PDs. As discussed in Section 3.1.1, a loan portfolio is segmented into PD bands or rating grades. According to Basel II (BCBS, 2005a, para. 501), lenders must calibrate the PD of each rating grade to ensure that the actual default rates are within the expected range. The goal of PD calibration is to determine whether the size of the difference between the estimated PDs and the observed default rates is significant (Crook *et al.*, 2007). Such a procedure is commonly referred to as *backtesting*.

The appropriate Basel II body [Validation Group of the BIS Research Task Force (BCBS:VG, 2005) (BISVG)] considered the Binomial test (see Siegel, 1956), the Chi-square test, and the Normal test (or Z test) (see Sprinthall & Fisk, 1990) as a means of testing for significance between the PDs. However, the power of the tests were found only to be moderate and the BISVG concluded that [(BCBS:VG, 2005) in Crook *et al.* (2007)] “*at present no really powerful tests of adequate calibration are currently available*”, and called for more research. The response to this call, with respect to LDPs, is described later in Section 4.1.1.

3.3 Conclusion

Credit scoring is used by banks to rank and assess the risk of loss arising due to any real or perceived change in customers’ ability and willingness to repay a financial obligation. Initially, the credit scoring of customers for retail loans was performed as

a subjective exercise by a bank manager. The 5Cs were used as a subjective guideline in determining whether or not to grant credit. In the USA, government policy and legislation (e.g. ECOA) on the access to homeownership and credit coincided with technological advances, all of which helped to accelerate the demand and adoption of objective credit scoring systems which assess lending risks based on empirical evidence. Prior to 1980, modern credit scoring was an American preserve, however other developed nations soon began to adopt the more advanced data-driven credit scoring systems (Anderson, 2007).

The correct functioning and refinement of credit scoring systems is an obvious topic of interest to banks, customers, and regulators alike - as recent events in the world economy have demonstrated. For example, under the Basel II Capital Accord, banks are now required to provide the relevant regulatory authorities with an accurate estimate of customers' probability of default (PD) as part of estimating banks' minimum capital requirements.

A scorecard is a numerical scale used to assign points to customer characteristics' in order to derive a numeric value which represents the risk that a customer, relative to other customers, will default on their financial obligation. The process to develop a scorecard consists of three main stages: (i) dataset construction; (ii) modelling; and (iii) documentation. This chapter has described a wide range of techniques and approaches used during the dataset construction and modelling stages.

Scorecard development is a detailed process that requires attention to many facets. Demographic trends and economic events can create scenarios for which the standard, accepted scorecard development techniques are unsuited. One such challenge arises when too few defaulters occur in the sample population, presenting

difficulties in constructing a robust and reliable scorecard. Another such challenge occurs in behavioural scoring when scorecard builders must decide on how to define a default and what length of time (i.e. the number of months) to base customer behaviour on. Finally, for many academics obtaining actual credit scoring data is a difficult task on account of data privacy laws and commercial sensitivities. By using artificial data academics can overcome these restrictions and create specific conditions under which to investigate specific problems. The next chapter presents and describes these specific problems in detail and explains how they impact the scorecard development process.

CHAPTER 4

Credit Scoring Challenges

In Chapter 3 the scorecard development process was described, with particular emphasis placed on tasks performed during the dataset construction and modelling stages. The purpose of this chapter is to describe in detail a number of specific challenges and problems scorecard builders encounter during the aforementioned stages. In particular, this chapter presents a review of the literature in respect to low-default portfolios (LDPs), behavioural scoring, and artificial data, which are the challenges that are the focus of this thesis.

Section 4.1 provides a review of the literature on LDPs. First, the implications of Basel II regulation with respect to LDPs is highlighted. Next, a review of existing studies into LDPs is provided. The applicability of certain classification techniques, namely supervised learning and one-class classification, to LDPs is then considered. Section 4.1 concludes with a proposal for a set of experiments to examine the limits of this applicability.

The second topic addressed by this work is behavioural scoring. The literature on behavioural scoring is reviewed in Section 4.2. Existing approaches to behavioural scoring are identified along with the typical features which constitute a behavioural scoring dataset. Section 4.2 concludes with the outline for an empirical study which investigates the main issues affecting the construction of behavioural scoring models.

The third and final topic examined in this work is the generation of artificial credit scoring data. Section 4.3 reviews approaches to generating artificial data, and motivates the need for artificial data by identifying inadequacies with two popular credit scoring datasets used in academia. In addition, the lack of data sharing amongst academics is highlighted and discussed. The merits and pitfalls of using artificial data are then reviewed. Section 4.3 concludes with an outline of the research conducted on artificial data in the remainder of this thesis.

In this work, we consider imbalanced credit scoring datasets to exhibit absolute rarity.

4.1 The Low-Default Portfolio Problem

At certain stages of an economic cycle the number of defaulters can be very low, which complicates the modelling process. Section 2.3 described how the performance of standard supervised classification techniques deteriorates in the presence of *imbalanced data*. Imbalanced data refers to a situation where one class is under-represented compared to the other class. In credit scoring imbalanced data is common due to the usual absence of defaulters and this is known as the *low-default portfolio problem*. In the context of class imbalance, low-default portfolios are con-

sidered as a case of absolute rarity.

To use the Basel II *internal ratings-based* (IRB) approach to regulatory capital, lenders must be able to build models that are validated to have consistent and accurate predictive capacity (BCBS, 2005a, Paragraph 500). This has raised concern in the financial industry that lenders with low-default portfolios may be excluded from the IRB approach due to inability to build and validate accurate models (BBA, 2004). As a consequence such institutions would be forced to use simpler approaches requiring greater amounts of regulatory capital.

4.1.1 Calibration of Low-Default Portfolios

Many of the papers addressing the LDP problem do not investigate the issue of comparing the predictive performance of classification models through out-of-sample testing. Focus is instead given to the application of various statistical techniques that attempt to bolster the information generated by the monotonic ordering of the portfolio or by the small number of defaults in the portfolio. Such papers are concerned with the accurate model validation of LDPs.

Christensen *et al.* (2004) reported on confidence sets for PD estimates by using a parametric bootstrap. Similarly, Hanson & Schuermann (2006) also employ bootstrap approaches to derive confidence intervals around estimates of default frequencies, however this approach requires a certain minimum number of defaults in at least some rating grades (Pluto & Tasche, 2011).

Pluto & Tasche (2006) [and (Pluto & Tasche, 2011)] address the LDP problem by proposing a method based on the “*most prudent estimation principle*”, which employs the idea of confidence intervals and uses an appropriate upper confidence

bound as a conservative default probability estimator (Orth, 2011). This approach relies on the assumption that the ordinal ranking of the borrowers, who are split into grades of decreasing credit-worthiness, is correct. Forrest (2005) adopted a similar method to Pluto & Tasche (2006), but in contrast this method is based on the likelihood approach by working in multiple dimensions, where each dimension corresponds to a rating grade and each point represents a possible choice of grade-level PDs. Benjamin *et al.* (2006) outline an approach to generating conservative estimates of LDPs using a look-up table, from which a look-up PD is calculated and compared to the weighted average PD of a firm's portfolio.

Other authors have examined the use of Bayesian methods for the PD estimation of LDPs. The incorporation of prior information can be particularly useful in small samples of data that provide only limited information on the parameter of interest (Orth, 2011). It is possible to use such prior information by specifying a prior distribution for the parameters of interest (i.e. PD). There are a number of different ways this can be achieved. Kiefer (2009) combines expert opinion, incorporated in the form of a probability distribution, with a Bayesian approach. This approach dispenses with the choice of a confidence level and instead relies on the subjective opinions of an expert trained in working with probabilities. This is a time-consuming process and may not satisfy regulators, particularly when an incentive exists for providing a less than conservative estimate of the priors (Orth, 2011). Dwyer (2007) also employs a modified Bayesian approach for validating the accuracy of the forecasted PD estimates, but without the use of expert information. Similarly, (Tasche, 2012) uses uninformed priors. Stefanescu *et al.* (2009) has also examined model calibration from historical rating transition data using a Bayesian

hierarchical framework. van der Burgt (2008) proposes an approach that is based on fitting the cumulative accuracy profile (CAP or Lorentz curve) to a concave function. Other notable works include Orth (2011) who use an empirical Bayes approach (see Carlin & Louis, 2008) and argues that the prior information can be obtained from additional datasets that supplement the original dataset. For example, a bank may hold an assortment of retail loan portfolios which can be used by the empirical Bayes approach to estimate the PD for each particular portfolio. However, this approach is probably more suited to sovereign bonds as supplementary data from ratings agencies (e.g. Standard & Poor's) can be utilised.

4.1.2 Modelling Low-Default Portfolios

The question of which classification technique to select for credit scoring remains a complex and challenging problem. Baesens *et al.* (2003) highlight the confusion resulting from comparing conflicting studies. Some studies may recommend one particular classification algorithm over another, whilst other studies recommend the opposite. Furthermore, many of these studies evaluate a limited number of classification techniques, restricted to a small number of credit scoring datasets. To compound this, many of the datasets are not publicly available, thus curtailing reproducibility and verifiability. Another problem is authors' expertise in their own method and failure to undertake a corresponding effort with existing methods (Michie *et al.*, 1994). Indeed, Thomas (2009b) highlights that studies which have endeavoured to avoid the aforementioned problems (Baesens *et al.*, 2003; Xiao *et al.*, 2006) have reported that the differences between the performance of classification techniques were small and regularly not statistically significant. Great care and con-

sideration was taken to avoid these issues in this thesis, details of which are given in Section 5.1.

Overall, the two main technical challenges presented by LDPs are: (i) estimating an accurate PD when no historical defaults are available; and (ii) assessing a model's predictive performance (Stefanescu *et al.*, 2009). Both of these issues arise not only during the validation of the model, but also prior to this, during the construction of the model. In many of the works addressing the LDP problem, the construction of the model is dependent on: (i) making assumptions about the ordering of the data; (ii) incorporating expert opinion; or (iii) the availability of a certain number of historical defaults generated either artificially or occurring in reality.

Given any of these dependencies, the models constructed are typically either: (i) statistical models constructed from a representative pool of data; or (ii) expert systems (or knowledge-based approaches) whose parameters are determined by financial experts. van Gestel & Baesens (2009) highlight several experimental studies from various domains outside of credit scoring which conclude that quantitative statistical models outperform human experts (e.g. Meehl, 1955). This is not to say that certain knowledge-based implementations, such as the BVR-I rating system used by the Federal Association of German Cooperative Banks (see OeNB/FMA, 2004), cannot be successfully utilised to achieve good predictive ability among loan applicants (Tang & Chi, 2005). Indeed, an advantage of such approaches is the ability to generate explanatory models which provide the expert with an explanation as to why a certain credit applicant is accepted or rejected (Hoffmann *et al.*, 2007). However, such systems are beyond the scope of this work in which we focus on quantitative approaches.

Much research has been conducted on adapting classification techniques to construct credit scoring models [e.g. logistic regression (Westgaard & Van der Wijst, 2001), neural networks (West, 2000)]. As a further example, a non-exhaustive list of such studies is available in Brown (2012). It is possible to combine many of these classification techniques to create an ensemble classification technique. Much of this research is performed on the basis that the constructed credit scoring models use datasets containing a representative number of historical defaults.

There is a paucity of studies in the literature assessing the LDP problem. One study by Brown & Mues (2012), conducts a comparison of several classification techniques on a range of credit scoring datasets with varying levels of class imbalance. Five real-life credit datasets are used, and for each dataset a further 8 datasets were created with good:bad ratios of: (i) 70:30; (ii) 75:25; (iii) 80:20; (iv) 85:15; (v) 90:10; (vi) 95:5; (vii) 97.5:2.5; and (viii) 99:1. The good:bad ratios were achieved by undersampling the majority class as required to achieve each ratio. The classification techniques used included linear discriminant analysis, quadratic discriminant analysis, logistic regression, least square support vector machines (linear kernel) (Suykens & Vandewalle, 1999), neural networks (multi-layer perceptron), C4.5 decision trees, the k -nearest neighbours algorithm (k -NN), random forests (Friedman, 2001, 2002), and gradient boosting (see Breiman, 2001). The performance of these techniques were assessed using the AUC, with Friedman's test and Nemenyi's *post hoc* tests applied to determine statistically significant differences between the average ranked performances of the AUCs. The study reported that at extreme levels of class imbalance the more complex techniques, gradient boosting and random forest classifiers, yielded a “*very good performance*”. However, the linear discriminant analysis and

logistic regression classification techniques, gave results that were reasonably competitive even at levels of high class imbalance.

As highlighted previously in Section 2.4, the training data used by one-class classification techniques consists of labelled examples for the target class only, as non-target class examples are too expensive to acquire or too rare to characterise. One-class classification techniques have already been successfully applied to a wide range of real-world problems, e.g. fault detection. To the best of our knowledge a benchmarking study of the performance of one-class classification techniques on low-default portfolios has not been described in the literature. The most closely related work is Juszczak *et al.* (2008), which describes a comparison of one- and two-class classification algorithms used for detecting fraudulent plastic card transactions. The results of that study found that two-class classifiers will outperform one-class classifiers - provided that the training and test objects are from the same distribution. Plastic card fraud detection is also examined by Krivko (2010) who provide a framework for combining one- and two-class classifiers to identify fraudulent activity on debit card transaction data.

Basel II regulation has established the LDP problem as an outstanding issue in credit scoring. Supervised classifiers trained to address such problems typically under-perform as the data on which they are trained is not representative of the concept to be learned. Further investigation is necessary to accurately ascertain the possibilities and the limits of using OCC techniques to address the LDP problem.

4.1.3 Low-Default Portfolios: Thesis Research

One possible approach to addressing the low-default portfolio problem is the use of *one-class classification* (OCC) algorithms such as the ones described in Section 2.4. As outlined above, OCC (also known as *outlier detection*) has attracted much attention in the data mining community (Chawla *et al.*, 2004). It is a recognition-based methodology that draws from a single class of examples to identify the *normal* or expected behaviour of a concept. This is in contrast to standard supervised classification techniques that use a discrimination-based methodology to distinguish between examples of different classes.

In Chapter 5 we compare OCC methods with more common two-class classification approaches on a number of credit scoring datasets, over a range of class imbalance ratios. As a means for handling imbalanced data we oversample the minority class along with adjusting the threshold value on classifier output. The purpose of this evaluation is to determine whether or not the performance of OCC methods warrants their inclusion as an approach to addressing the LDP problem. To the best of our knowledge, no attempt has been made to examine OCC as a solution to the LDP problem before.

4.2 Behavioural Scoring

Behavioural scoring, is used after credit has been granted and estimates an existing customer's likelihood of default in a given time period. Behavioural scoring allows lenders to regularly monitor customers and help coordinate customer-level decision

making. The data used for model fitting for this task is based on the customers' loan repayment performance and also their good/bad status at some later date. To be profitable a bank must accurately predict customers' likelihood of default over different time horizons (1 month, 3 months, 6 months, etc.). Customers with a high risk of default can then be flagged allowing the bank to take appropriate action to protect or limit itself from losses.

Behavioural scoring is used by organisations to guide lending decisions for customers in: credit limit management strategies; managing debt collection and recovery; retaining future profitable customers; predicting accounts likely to close or settle early; offering new financial products; offering new interest rates; managing dormant accounts; optimising telemarketing operations; and predicting fraudulent activity (Hand & Henley, 1997; Malik & Thomas, 2009; McDonald *et al.*, 2012; McNab & Wynn, 2000; Sarlja *et al.*, 2009).

The financial circumstances of a customer are likely to change over time, and as such, they are continuously monitored and managed. The first behavioural scoring system to predict credit risk of existing customers was developed by Fair Isaac Inc. for Wells Fargo in 1975¹. Behavioural scorecards have since evolved to influence decisions across the entire credit cycle. For example, *usage scorecards* for credit card products attempt to predict future levels of activity to assist in retention and incentive strategies. *Account management scorecards* are used during the lifetime of an account by lenders to predict the risk of default at a given point in time (e.g. every month, quarter, year). This allows the lender to set loan limits on top-up loan decisions and take appropriate measures to contain bad, loss-making

¹<http://www.fico.com/en/Company/Pages/history.aspx>

accounts. Such information is also valuable to lenders' marketing departments when selecting profitable customers for additional products or deciding to what extent to incentivise increased account usage. A detailed list of the different types of behavioural scorecards is provided in McNab & Wynn (2000).

4.2.1 Behavioural Scoring: Approaches

Broadly speaking, there are two approaches to behavioural scoring: techniques that use static characteristics about the customer's past performance; and techniques which incorporate dynamic aspects. Thomas *et al.* (2001b) survey the approaches and objectives of behavioural scoring, with particular focus on procedures that incorporate dynamic aspects of customer behaviour, e.g. Markov models (see Malik & Thomas, 2012). This thesis does not examine behavioural scoring techniques which incorporate dynamic aspects of customer behaviour.

Figure 4.1 illustrates the longitudinal aspect to the data used in behavioural scoring. A sample of customers is selected so that their repayment behaviour either side of an arbitrarily chosen *observation point* is available. The period before the observation point is often termed the *performance window*. Data on the customers' performance during this time is structured into features which are used by the behavioural scoring system to distinguish between customers' likely to repay their loan and those likely to default on their financial obligation.

The data used in the performance window is derived from the banks' own internal databases and external data sources such as credit bureaus. The data describes customers' demographic (e.g. date of birth, address), transactional (e.g. purchase history), and performance (e.g. arrears) features. Based on the work of McNab

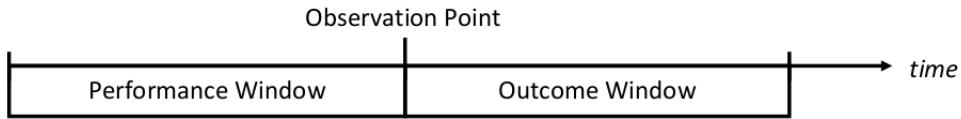


Figure 4.1: Behavioural scoring performance window and outcome window.

& Wynn (2000), Table 4.1 lists the sources of typical behavioural scoring features. The appropriateness of the features will vary depending on the behavioural scoring system. For example, features from the *promotions history* are unsuitable for behavioural scoring of retail loans with fixed long-term repayment periods.

Table 4.1: Behavioural scoring data sources and associated feature examples (McNab & Wynn, 2000).

<i>Data Source</i>	<i>Feature Example</i>
Delinquency history	Ever in arrears Maximum arrears level
Usage history	Balance-to-limit ratio Balance trends
Static information	Customer age Application score
Payment/purchase history	Purchase frequency Type of retail goods purchased
Collections activity	Outcomes Contact frequency
Revolving credit transactions	Number Type (retail/cash)
Customer service contacts	Inbound contact Outbound contact
Promotions history	Number of offers Outcome of offers
Bureau data	Generic scores Shared account information

The period after the observation point is known as the *outcome window*. The purpose of the outcome window is to classify borrowers into distinct populations (i.e. good and bad) based on their level of arrears (Mays, 2004). The correct classification

is unavailable at the time when the performance window data arrive and the prediction is needed. However, in order to be able to train the behavioural scoring model, we assume that the correct classification is available during the training phase. Selecting an appropriately sized outcome period requires careful consideration. As this period of time is used to classify customers, a comprehension of economic conditions, operational policies, and borrower volatility is necessary. If this period of time occurs during favourable economic conditions, then the performance of the scoring model may degrade if the reverse is true. Financial institutions need to consider the effects on customer behaviour caused by adopting certain operational policies, or *policy bias* Thomas (2009a). For example, an early intervention policy may reduce the incidence of customers missing further loan repayments and subsequently being classed as bad. Finally, the outcome period should be sufficiently sized so as to capture a representative sample of bards with which to build a stable behavioural scoring model.

Typically, the same techniques used in application scoring are used in behavioural scoring to classify customers into one of two categories: good and bad (Thomas *et al.*, 2001b). Behavioural scorecard modellers encounter many of the same scorecard development and implementation issues as with application scoring, such as: identifying and adjusting for different segments of the population (see Bijak & Thomas, 2012), ensuring the optimal correlation between features (see Tsai, 2009), handling class imbalance (see Burez & Van den Poel, 2009), and identifying the correct sample size (see Crone & Finlay, 2012).

To build behavioural scoring models practitioners must make decisions on a number of important parameters. This involves asking pertinent questions such as:

The extensiveness of the historical data with which to model customer performance?

How far forward into the future to make reliable predictions? What defines a loan defaulter? The credit scoring literature does not contain strong recommendations on how to answer these questions.

To the best of our knowledge, to date very little empirical research has been published in the literature investigating the effects of different sized time horizons on classifier performance. Much of the recent research for determining appropriate time periods in behavioural scoring is conducted in the context of assessing the applicability of a particular duration model, survival analysis (see Andreeva, 2005), as a method of identifying loan defaulters. With duration models, the focus is not whether an applicant will default, but if they default when will this occur (Banasik *et al.*, 1999). Duration models are not within the scope of this work, and are not examined in this thesis.

Finally, Section 3.2.1.4 described how financial institutions can classify customers as either good or bad using either a *current status* or *worst status* label definition approach. To date, no study comparing both these approaches has appeared in the literature.

4.2.2 Behavioural Scoring: Thesis Research

This thesis investigates some of the main issues affecting the construction of behavioural scoring models by examining the performance of retail loans issued by the main Irish banks in 2003 and 2004. The findings reported in this thesis are based on real-world data from a credit bureau.

First, we compare the accuracy of scoring models that are built using different

historical durations of customer repayment behaviour (6-months, 12-months, and 18-months). Next, we quantify the differences between varying outcome periods from which a customer’s class label is predicted (3-months, 6-months, 12-months, 18-months, and 24-months). Finally, we demonstrate differences between alternative approaches used to assign customers’ class label (*current status* or *worst status*).

4.3 Artificial Data

In credit scoring, over the last decade numerous studies examining the performance of various models used to construct credit scorecards have been produced, (e.g. Baesens *et al.*, 2003; Chen *et al.*, 2011; West, 2000). A concern is that the data used in these studies originates from two sources: (i) the Australian and German datasets which are publicly available from the University of California Irvine (UCI) Machine Learning Repository (Asuncion & Newman, 2007); and (ii) private datasets obtained from financial institutions.

The UCI repository serves several important functions (Salzberg, 1997). The repository allows published results to be checked and, through comparison with existing results, allows researchers assess the plausibility of a new algorithm. A number of researchers (Martens *et al.*, 2011; Salzberg, 1997; Soares, 2003), however, caution against over-reliance on the UCI repository as a source of research problems. The repository is cited as a potential source of over-fitting, as researchers’ familiarity with datasets from the repository may influence them to design algorithms that are tuned to the datasets (Salzberg, 1997). As a result researchers often ignore the problem of trying to understand under which conditions an algorithm works

best (Soares, 2003). It is beneficial that researchers do not over-rely on the UCI repository, preferably multiple data sources should always be used.

Another issue that has been raised with datasets from the UCI repository is that the datasets are not truly reflective of the real-world and only capture a small subset of all of the situations that can arise in real-world scenarios (Drummond & Holte, 2005a; Saitta & Neri, 1998). The Australian and German credit application datasets, for example, contain very different class distributions and the overall size of the datasets is not representative of datasets that occur in modern practice. Such differences are likely to be an artefact of how the datasets were constructed, which in turn raises questions about how the data was collected (Drummond & Japkowicz, 2010). Given the different class distributions, the assumption that the sampling of these datasets is random needs to be treated with caution. The inclusion of certain features also raises questions about the current relevancy of the UCI data. For example, in an age of the ubiquitous mobile phone, the use of a *telephone* feature in the German dataset is questionable. Consequentially, one should be careful not to derive too much from experimental results using these datasets alone.

It is desirable to use data derived from more diverse sources. Researchers could achieve this diversity by using multiple real-world datasets obtained directly from a financial institution or via another researcher. However, for researchers lacking the necessary resources, obtaining real-world data is a source of great frustration. Fischer & Zigmond (2010) describe a number of factors impeding the sharing of data within academia, which are reiterated below.

Negative Career Impact The need to publish is important to a researcher's career and datasets may be part of a long-term endeavour from which an individ-

ual could generate multiple publications. If a researcher is required to share data after their first publication the opportunity to generate further publications may be reduced if a better funded and resourced research group obtains the data.

Limited Resources Sharing data may require extra resources to convert it into an accessible format for other researchers. This reduces the time and money available to the originator to pursue their own research activities. Certain datasets may also require updating and maintenance, and once a researcher has completed their work it may be no longer feasible to store the data.

Property Rights and Legal Issues Legal and commercial reasons may prohibit the researcher from sharing data. Customer confidentiality, for example, is of the utmost importance for any financial institution. While with a single anonymised dataset it may not be possible to identify a particular individual, customer identity may be compromised through the combination of multiple datasets.

The authors are of the opinion that the above barriers will remain in place for the foreseeable future. Principally this is due to a lack of incentives for the originator to share data, and the overly stringent requirements of data protection laws (see Bergkamp, 2002). Without the provision of publicly available datasets, credit scoring will remain closed to the wider data mining community.

In order to overcome the aforementioned difficulties we highlight the benefits of using artificial data. An advantage artificial data has over real-life data is the flexibility afforded to the manipulation of various parameters used in the evaluation process. Using artificial data, the researcher has the capability to (Malin & Schlapp, 1980): (i) include as many or as few data samples as they choose; (ii) specify the precise distribution of the data present; (iii) include noise with a known standard

deviation; and (iv) test the effects of other variations. Using this approach, the researcher can design specific experiments aimed at evaluating the performance of algorithms under particular conditions of interest in a relatively precise manner (Scott & Wilkins, 1999).

It is important to recognise, however, that the inherent unpredictability of real-world data (e.g. natural disasters, unforeseen changes in personal circumstances) cannot be replicated using artificial data. One cause of this unpredictability is the structural complexities arising from external and uncaptured circumstances (Scott & Wilkins, 1999). This cannot be replicated as structural regularity must be imposed on artificial data in the form of some fixed distributional model (Japkowicz & Shah, 2011). Furthermore, even though unintended, because of the way it is generated artificial data can be biased towards a particular classification technique that is capable of modelling the data more precisely than others. Caution must be exercised when interpreting findings obtained using artificial data as it is analogous to "*laboratory conditions*" and may not necessarily translate to real-world conditions. For these reasons, one should judiciously select the research questions that artificial data will be used to answer (Japkowicz & Shah, 2011). Despite these concerns, using artificial data allows a researcher to clearly conceptualise a problem. This allows the researcher to establish their understanding of the basic assumptions of the problem along with the imposed constraints. Provided that the beliefs and assumptions used to generate the data are valid the researcher can then proceed with evaluating and interpreting the behaviour of existing and novel approaches used on the same or related problems.

There are many challenges in the credit scoring domain for which artificial data

can be used as part of the investigation. One such area of interest addressed in Section 7.2 illustrates how changes to the underlying credit scorecard population affect the predictive accuracy of a classifier. Another example worthy of consideration, though not explicitly examined in this thesis, involves determining to what extent the quality of the data affects credit scorecard construction. Using artificial data it is possible to design experiments with which to examine the data quality characteristics previously specified in Section 3.2.1.1. It must be stressed, however, that any conclusion found using artificial data must be verified using real data.

4.3.1 Artificial Data: Previous Work

In credit scoring, researchers typically experiment with artificial data in order to demonstrate the feasibility of some proposed classification algorithm. A straightforward approach, as used by Hand & Adams (2000), is to generate data according to some p -dimensional multivariate normal distribution with a specified mean vector μ and covariance matrix Σ for each class. An even simpler approach is to use two univariate Gaussian distributions as this allows for visualisation of the model (see Hoffmann *et al.*, 2007; Kelly *et al.*, 1999). Publicly available artificial datasets such as Ripley's dataset (Ripley, 1994) have also been used (see Martens *et al.*, 2007).

Whilst such artificially generated data is useful for demonstrative purposes, the findings may not necessarily translate to real-world conditions. Generally, there are two approaches to address the lack of availability of real-world data:

- Use existing real-world data as a “seed” with which to generate artificial data.
- Generate artificial data without using any real-world data.

An example of the first approach is the Synthetic Minority Oversampling Technique (SMOTE) (Chawla *et al.*, 2002), which is an oversampling approach where the minority class is oversampled by creating synthetic examples rather than by oversampling with replacement.

As an example to the second approach, Andersson *et al.* (2011) highlight how regulators in the USA use artificially generated credit scores which are correlated with systematic factors (e.g. unemployment rate) to validate credit scoring models. Outside of credit scoring, many specialised dataset generators have been described in the literature. For example, Scott & Wilkins (1999) describe two artificial data generators, one based on the multivariate normal distribution and the other inspired by fractal techniques for synthesising artificial landscapes. Other examples include an IBM dataset generator (Srikant, 1994) which simulates a retail environment and produces market baskets of goods; and *celsim* (Myers, 1999) used in the genome assembly process by generating a user described DNA sequence with a variety of repeat structures along with polymorphic variants.

Alaiz-Rodríguez & Japkowicz (2008) simulate a medical domain that states the prognosis of a patient a month after being diagnosed with influenza. Each patient is described using a number characteristics: (i) patient age; (ii) influenza severity; (iii) patient's general health; and (iv) patient's social status. Three of the characteristics (age, influenza severity, and social status) are completely independent and one characteristic (general health) is dependent on two other characteristics (age, and social status). The prognosis class depends on all four characteristics and the data is generated based on user defined prior probabilities for each characteristic. By manipulating the prior probabilities for each characteristic the user can simu-

late various scenarios (e.g. an increasingly virulent influenza outbreak, developing population, or poorer population).

A number of general frameworks for generating data also exist (Atzmueller *et al.*, 2006; Melli, 2007), however such general frameworks cannot replicate the rich complexity and intricacies of a specific domain, such as credit scoring. To the best of our knowledge no published framework exists for generating artificial credit scoring data.

4.3.2 Artificial Data: Thesis Research

In the present work, to help overcome the lack of data sharing in credit scoring we propose the design and development of a framework for generating artificial data. The main purpose of our artificial data framework is to provide researchers with a means of creating artificial (but suitably realistic) credit scoring datasets with which to assess the behaviour of classification techniques. Such datasets can enhance research in data mining and credit scoring (e.g. model development or assessing performance metrics). Furthermore, artificial data can help overcome limitations caused by both a deficient data sharing culture and a shortage of reliable real-world credit scoring datasets.

In domains where access to real-world data may simply be unattainable, and to overcome the aforementioned limitations currently associated with machine learning data repositories, we contend that the use of artificial data is acceptable. It should be stressed that the data must be generated in the correct manner and be sufficiently grounded in reality in order to avoid the danger of investigating imaginary problems.

4.4 Conclusion

This chapter has identified three separate, yet demanding, challenges encountered by both credit scoring researchers and practitioners. The first relates to low-default portfolios which arise from a shortage of loan defaulters. In this work we propose to evaluate the applicability of one-class classification techniques to the LDP problem. This work will also examine the effectiveness of oversampling the minority class along with adjusting the threshold value on classifier output as an approach to addressing class imbalance.

This chapter has also reviewed behavioural scoring and described some of the challenges encountered when constructing such models. This work will provide guidance, not readily available in the literature, on a number of important questions (for example, how to define a loan defaulter? Or what range of historical data to use when modelling customer performance?). These questions will be answered in this work with the construction of behavioural scoring models using credit bureau data from 2003 up to 2010. An empirical study will then quantify differences between the performance of the models using a collection of behavioural scoring datasets.

Finally, we have also described the impediments researchers encounter when attempting to obtain real-world credit risk data. Often legal requirements and commercial sensitivities prevent the sharing of data amongst the research community. It is our hope that this discussion will lead to greater understanding and awareness of the issue. Furthermore, we have outlined the benefits of sharing data amongst researchers to help impress upon the key stakeholders within financial institutions the potential rewards of data sharing within the credit risk community. In this work,

to help overcome the lack of data sharing in credit scoring, we propose a framework with which to generate artificial data suitable for use in credit scoring.

CHAPTER 5

Using Semi-supervised Classifiers for Credit Scoring

As described in Chapter 2 (Section 2.3), class imbalance presents a problem to most supervised two-class classification algorithms as they assume a balanced distribution of the classes. In credit scoring, a particularly severe form of class imbalance is referred to as the low-default portfolio (LDP) problem. LDPs are characterised by an insufficient default history, such that the average observed default rates (i.e. the total number of defaults divided by the total number of loans for the entire portfolio) may be statistically unreliable estimators of default probabilities (Florez-Lopez, 2009).

This chapter investigates the suitability of oversampling as a solution to a form of class imbalance known as absolute rarity. In addition, this chapter evaluates what improvement in classification performance can be achieved by optimising the

threshold value on classifier output. The suitability of semi-supervised one-class classification algorithms as a solution to the low-default portfolio (LDP) problem is evaluated. In this chapter the performance of semi-supervised one-class classification algorithms is compared with the performance of supervised two-class classification algorithms. Assessment of the performance of one- and two-class classification algorithms using nine real-world banking datasets, which have been modified to replicate LDPs, is provided.

The comparative assessment of classification methods can be a subjective exercise. It is influenced, among other factors, by the expertise of the user with each of the methods used and the effort invested in refining and optimising each method (Hand & Zhou, 2009; Thomas, 2009b). We attempt to overcome this problem by restricting our study to a single application area (LDPs); by selecting appropriate performance measures (H measure and harmonic mean); and finally by using nine different datasets of varying size and dimension to capture as many as possible of the particular aspects of the LDP problem.

The remainder of this chapter is organised as follows. Section 5.1 describes the experimental methodology, and Section 5.2 presents experimental results. Section 5.3 discusses conclusions and directions for future work.

5.1 Evaluation Experiment

The aims of the evaluation described are to examine the effectiveness of oversampling and the use of one-class classification (OCC) in addressing the LDP problem. This is achieved by comparing the performance of one-class classifiers to that of the more

typical two-class classifiers. Furthermore, we investigate to what extent optimising the threshold value on classifier output yields an improvement in classification performance. To accomplish the above aims we adopt four separate approaches:

- i. Classifying an imbalanced dataset using a selection of two-class classifiers.
- ii. Oversampling the minority class of the dataset and employing a selection of two-class classifiers.
- iii. Removing the minority class completely and using OCC.
- iv. Repeating (i) - (iii) whilst optimising the threshold value on classifier output for each of the one- and two-class classifiers.

The first two approaches compare various two-class classifiers on credit scoring datasets with different degrees of class imbalance. Both approaches illustrate the adverse consequences of class imbalance. Based on the results of (i) and (ii) the best performing two-class classifier is then used in approach (iii) where its performance is compared to a selection of one-class classifiers on the same datasets but with a greater degree of class imbalance. This process is then repeated using an optimised threshold value whenever classification is performed. The remainder of this section describes the datasets, performance measures and methodology used.

5.1.1 Data

The premise for selecting the datasets used in this evaluation is that (i) the datasets have been used in previous credit scoring studies; and (ii) it must be possible for other academic researchers to access the datasets so as to ensure the replicability

of experimental results. A total of nine datasets that matched these criteria were identified. Details and contact information to obtain the datasets are included in Appendix C. The datasets have previously been used in studies concerning corporate bankruptcy prediction, assessing individual applicants for revolving credit products, and assessing retail loan applicants [(e.g. see Hand, 2009; Možina *et al.*, 2007; Tsai, 2009; West, 2000; Xie *et al.*, 2009)]. The characteristics of the datasets used are presented in Table 5.1. The Australia, German and Japan credit datasets are publicly available at the UCI Machine Learning repository¹. The Japan credit screening dataset has been commonly mistaken for the Australia dataset, for example, by Tsai & Wu (2008) and Nanni & Lumini (2009). This incorrect version of the Japan dataset contains the same distributions and feature values (albeit with different attribute labels) as the Australia dataset. We use the correct version, a credit screening dataset, which stores the data in a LISP file format. The Iran dataset is an updated version of a dataset that appears in Sabzevari *et al.* (2007). It consists of corporate client data from a small private bank in Iran. The Poland dataset contains bankruptcy information of Polish companies recorded over a two-year period (Pietruszkiewicz, 2008). The Spain dataset compiled by Dionne *et al.* (1996) comes from a large Spanish bank and details personal loan applicants. The Thomas dataset is a CD ROM accessory of Thomas *et al.* (2002) describing applicants for a credit product. Two of the original fourteen features have been removed due to incomplete records. The Pacific-Asia Knowledge Discovery and Data Mining conference (PAKDD) dataset is a modified version of the PAKDD 2009 competition dataset. We removed redundant features and in order to reduce the size of

¹<http://archive.ics.uci.edu/ml/datasets.html>

the dataset we limit our selection of instances based on four different *phone code* feature values. The University College of San Diego (UCSD) set is also a modified competition dataset used in the 2007 University College of San Diego/Fair Issac Corporation (UCSD/FICO) data mining contest. We randomly undersampled both classes to reduce the size of the dataset and removed redundant identity features.

Table 5.1: Characteristics of the nine datasets used in the evaluation experiment. # Numeric refers to the number of continuous features and # Nominal refers to the number of categorical features.

	# Numeric	# Nominal	# Instances	# Good	# Bad	Good:Bad
Australia	6	8	690	307	383	44:56
German	7	13	1,000	700	300	70:30
Iran	19	2	413	332	81	80:20
Japan	5	5	125	85	40	68:32
PAKDD	6	10	1,764	1,404	360	80:20
Poland	30	0	240	128	112	53:47
Spain	1	17	2,446	2,110	336	86:14
Thomas	11	1	1,225	902	323	74:26
UCSD	32	6	5,397	2,684	2,713	50:50

All numerical attributes are normalised to values between 0 and 1 by applying min-max range normalisation. The sample sizes vary considerably from 125 to 5,397 instances. Typically, commercially used credit scoring models are usually constructed from an initial sample size of between 10,000 and 50,000 (Thomas, 2009b), which is reduced to a development sample of approximately 4,500 to 6,000 (consisting of between 1,500 and 2,000 each of good, bad, and declined applicants). However, the above datasets are all easily available in public literature whereas many other datasets used in credit scoring studies are privately held and cannot be shared amongst researchers. As per Keogh (2007), we believe that the irreproducibility of results caused by, amongst other things, the refusal to share data or to give parameter settings hinders the research process. To ensure reproducibility of the

contents of this chapter, we have provided access to all of the data and developed techniques used in this chapter. Indeed, for certain scientific disciplines it is not uncommon for journals and academic conferences to require public data deposition prior to publication. In keeping with best practices all of the datasets used in this chapter are publicly available¹.

5.1.2 Performance Measures

Two evaluation measures are used in this study: the harmonic mean (see Equation 2.25) and the H measure (see Section 2.5.1.3). The harmonic mean measures classification performance at a specific classification threshold, whereas the H measure assesses classifier performance over a distribution of costs. The harmonic mean of two numbers tends to be closer to the smaller of the two. This is an attractive feature for low-default portfolios, as such small values can often occur when measuring the classification accuracy of the minority class. The H measure is a relatively recent approach to classifier performance measurement, and may be considered as an alternative to the AUC and KS (see Hand, 2009). Demšar (2006) examined the problem of comparing classifiers on multiple datasets and recommends a set of simple and robust non-parametric tests for statistical comparisons of classifiers. Based on recommendations given by Demšar (2006), differences between the performance of various techniques were analysed with a Friedman test (Friedman, 1937) with *post hoc* pairwise comparisons performed with a Holm's procedure (Holm, 1979) (all tests for significance were at the 5% level).

¹Refer to Appendix C for details.

5.1.3 Methodology

To assess the effectiveness of oversampling and OCC, the distribution of the data must be altered to replicate LDPs. As the standard approach to assess credit scoring systems is to use a holdout test set, each dataset used was divided into three subsets:

(i) the training set (55%); (ii) the validation set (15%), and (iii) the test set (30%).

The training set and the validation set were used to train and tune the classifiers while the test set was used to assess their performance. This procedure was performed repeatedly over a number of turns. At the end of each turn the number of instances in the defaulter class of the training set was reduced by 10%. The model was then retrained and retuned using the training and the validation sets. To ensure that the appropriate classifier parameters were estimated, the class distribution of the validation set remained unmodified so as to correspond with that of the test set.

Figure 5.1 illustrates this process which we refer to as the *normal process*. It should be noted that for a real-world imbalanced dataset it is not common practise during the training phase to set aside a significant number of defaulters in a separate validation set. Typically, a process of k-fold cross validation is used where a small number of defaulters are present in the imbalanced datasets. In our approach, it must be acknowledged that the availability of *additional* defaulters (from the validation set) may potentially benefit two-class classifiers which are sensitive to tuning¹.

We conduct a second set of experiments on the same datasets whereby we oversample the number of instances from the defaulter class. This process was similar to the normal process except that after reducing the instances in the defaulter class of

¹Comment from external examiner

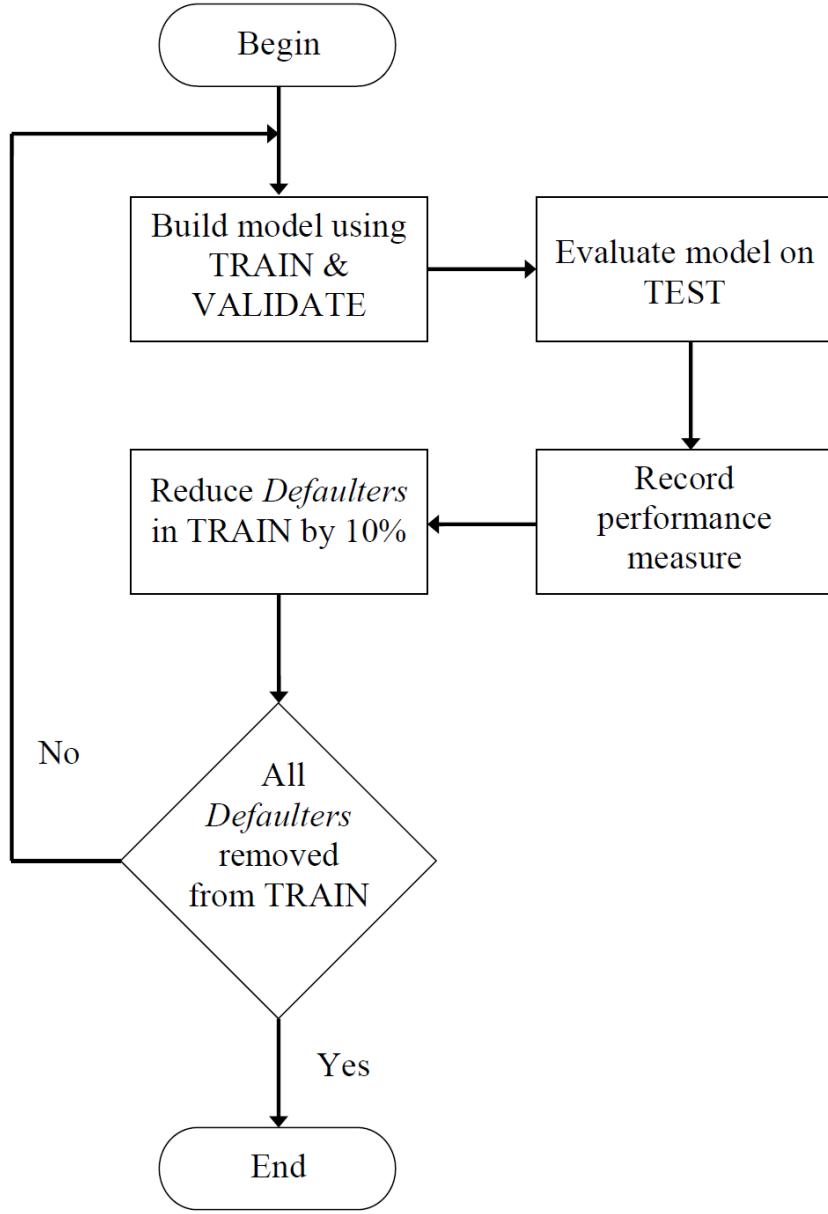


Figure 5.1: Normal process; training set - TRAIN, validation set - VALIDATE, test set - TEST.

the training set by 10% the remaining defaulter class instances were oversampled to produce a balanced training set. This oversampling occurs in the training data only. The validation set and the test set remain unchanged. We call this the *oversample process*, Figure 5.2 illustrates the procedure.

Finally, a third set of experiments using one-class classifiers is performed. When

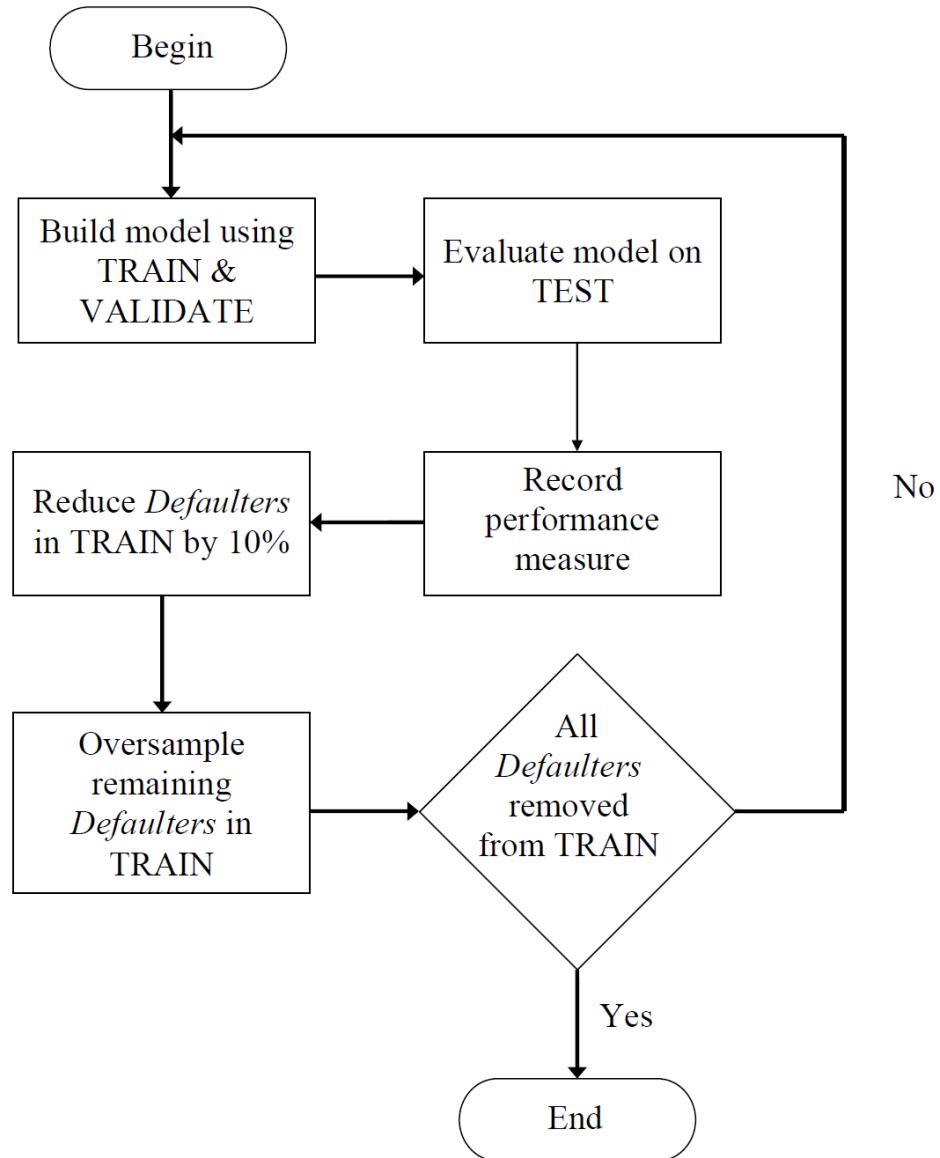


Figure 5.2: Oversample process; training set - TRAIN, validation set - VALIDATE, test set - TEST.

one-class classifiers are used we remove all the instances of the defaulter class from the training set, so that only instances from the non-defaulter class are used to build the model. Again, the validation set and testing set remain unchanged. We call this process the *one-class classification process* (OCC process).

In all three groups each experiment was conducted 10 times using different randomly selected training, test and validation set splits and the results reported are

averages of these 10 runs.

5.1.4 Classifier Tuning

The classification techniques used in this evaluation are previously described in Section 2.2 and Section 2.4. In our work, we exhaustively varied the parameter settings of each parameterised classifier in order to obtain the optimal values. The naïve Bayes (NB), linear Bayes normal (LDC) and Fisher’s linear discriminant analysis (LDA) supervised classifiers require no parameter tuning. For the quadratic Bayes normal (QDC) classifier, the regularisation parameters used to obtain the covariance matrix were optimised using the validation set. For the neural network (NN) the number of hidden layers was fixed at 1 and the number of units in the hidden layer matched the dimensionality of the input space, as per Piramuthu (1999), making the time consuming grid search procedure unnecessary. As a drawback, however, a relatively large number of hidden units may result in overfitting (Le Cun *et al.*, 1990). The k -nearest neighbour (k -NN) classifier uses $k = 10$ and Euclidean distance to determine the similarity between instances. For the logistic regression (LOG) classifier the number of cross validation iterations used during the maximum likelihood estimation method (see Section 2.2.2) to obtain the optimal feature class weights is optimised between 1 and 20 using the validation set. The Support Vector Machine (Lin SVM) classifier uses a linear kernel and the cost function parameter (i.e. regularisation parameter) is fixed at 0.5. When compared with an optimised cost function parameter obtained using a grid search process, a fixed cost function parameter provides stable results across the imbalance ratios without overfitting. The linear Bayes normal, Fisher’s linear discriminant analysis, neural network, and

quadratic Bayes normal supervised classifiers were implemented using PrTools (Duin *et al.*, 2008). The k -NN, logistic regression, and naïve Bayes supervised classifiers were implemented using the Weka (version 3.7.1) machine learning framework (Witten & Frank, 2005). The SVM classifier was implemented using LibSVM (Chang & Lin, 2001).

The one-class classifiers were implemented using the Matlab DDTools toolbox (Tax, 2009). For the Gaussian (Gauss) one-class classifier the regularisation added to the estimated covariance matrix is optimised using the validation set. For the mixture of Gaussians classifier (MOG), the number of clusters containing defaulters is optimised between 1 and 3 using the validation set. For each cluster the full covariance matrix was calculated. The regularisation for the covariance matrices was optimised using the validation set. For both the k -Means and k -NN classifiers k was set at 10. Both the Parzen and naïve Parzen (NParzen) used automated parameter settings. For the Auto-encoder (AE) the number of hidden layers was fixed at 1 (the default value). The number of hidden units was set to 5 (the default setting). With the Support Vector Data Description (SVDD), the parameter controlling the tightness of the boundary, σ , was optimised between 1 and 12 using the validation set. The range of values for σ were selected based on initial experimentation which showed that too high a value for σ resulted in overfitting.

The success of feature selection is very much dataset dependent (e.g. see Liu & Schumann (2005)), therefore the effects of feature selection techniques on the predictive performance of the described classification models are beyond the scope of this work. The next section will describe the results of this experimental process.

5.2 Results and Discussion

Figure 5.3, Figure 5.4, and Figure 5.5 illustrate the resulting H measure when eight two-class classifiers using the normal process and eight one-class classifiers using the OCC process were tested on the Australia, German, and Thomas datasets, respectively. The horizontal axis represents the percentage of defaulters present in the training dataset. The H measure is represented by the vertical axis. The two-class classifiers are identifiable by the deteriorating performance caused by the gradual removal of defaulters from the training dataset. As the number of non-defaulters used to train the one-class classifiers is fixed, the performance of the one-class classifiers remains static throughout. Figures C.1 to C.6 in Appendix C are included to demonstrate the performance of classifiers measured using the H measure under the normal process and the OCC process on the other 6 datasets. Figures C.30 to C.38 in Appendix C are also included to demonstrate the performance of classifiers measured using the AUC (see Section 2.5.1.2) under the normal process and the OCC process for all 9 datasets. The patterns present in Figure 5.3, Figure 5.4, and Figure 5.5 for the Australia, German, and Thomas datasets are broadly similar for all datasets.

Three separate segments have been highlighted in Figure 5.3, each representing a particular level of class imbalance (70:30, 80:20, and 90:10) at which we compare the performance of two-class classifiers. These three segments are selected as they represent a broad range of class imbalance. For some datasets (e.g. Spain) the initial level of imbalance only allows the two-class classifiers to be compared at class imbalances of 80:20 or 90:10.

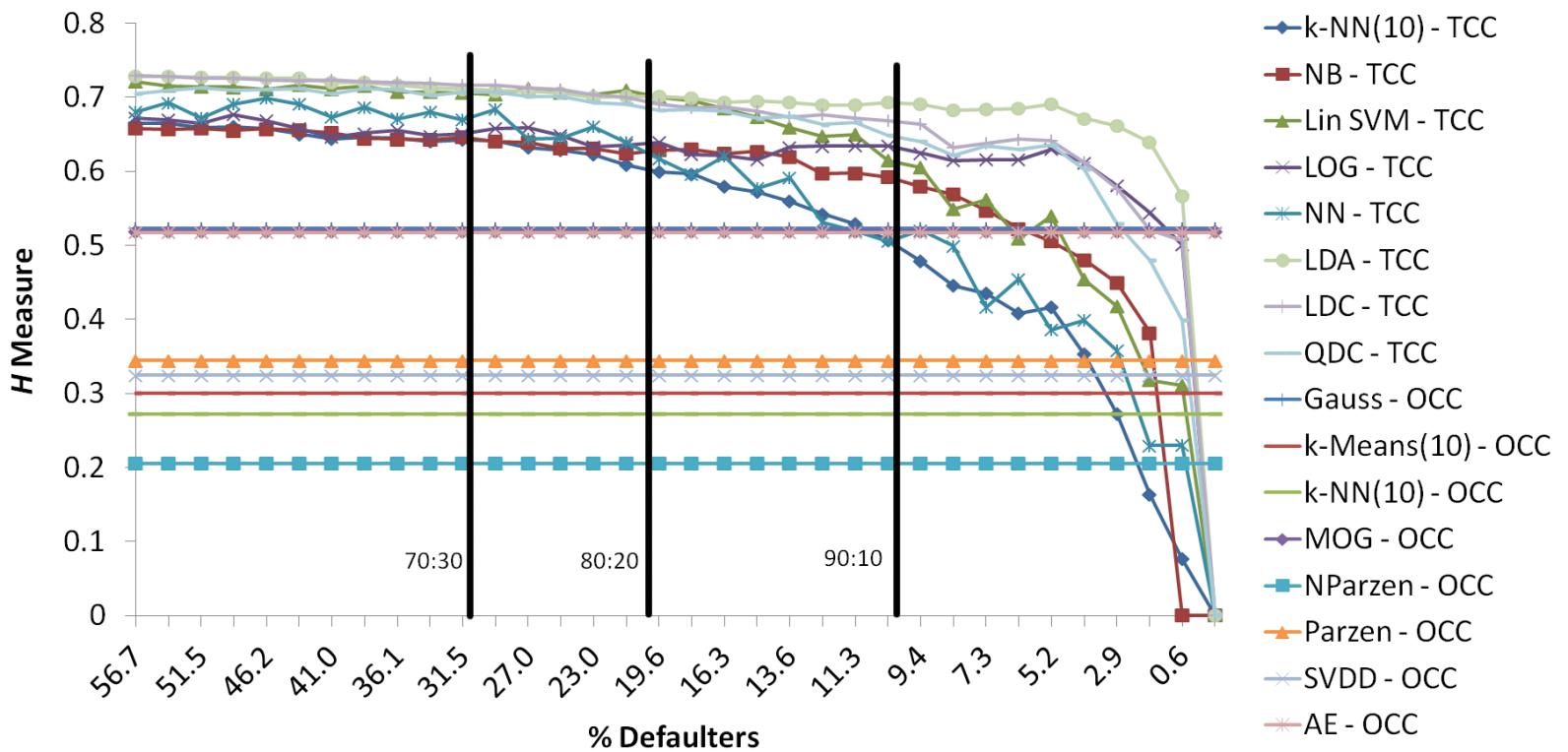


Figure 5.3: Australia: Normal process and one-class classification process test set H measure performance. Selected class imbalance ratios are also highlighted at 70:30, 80:20 and 90:10.

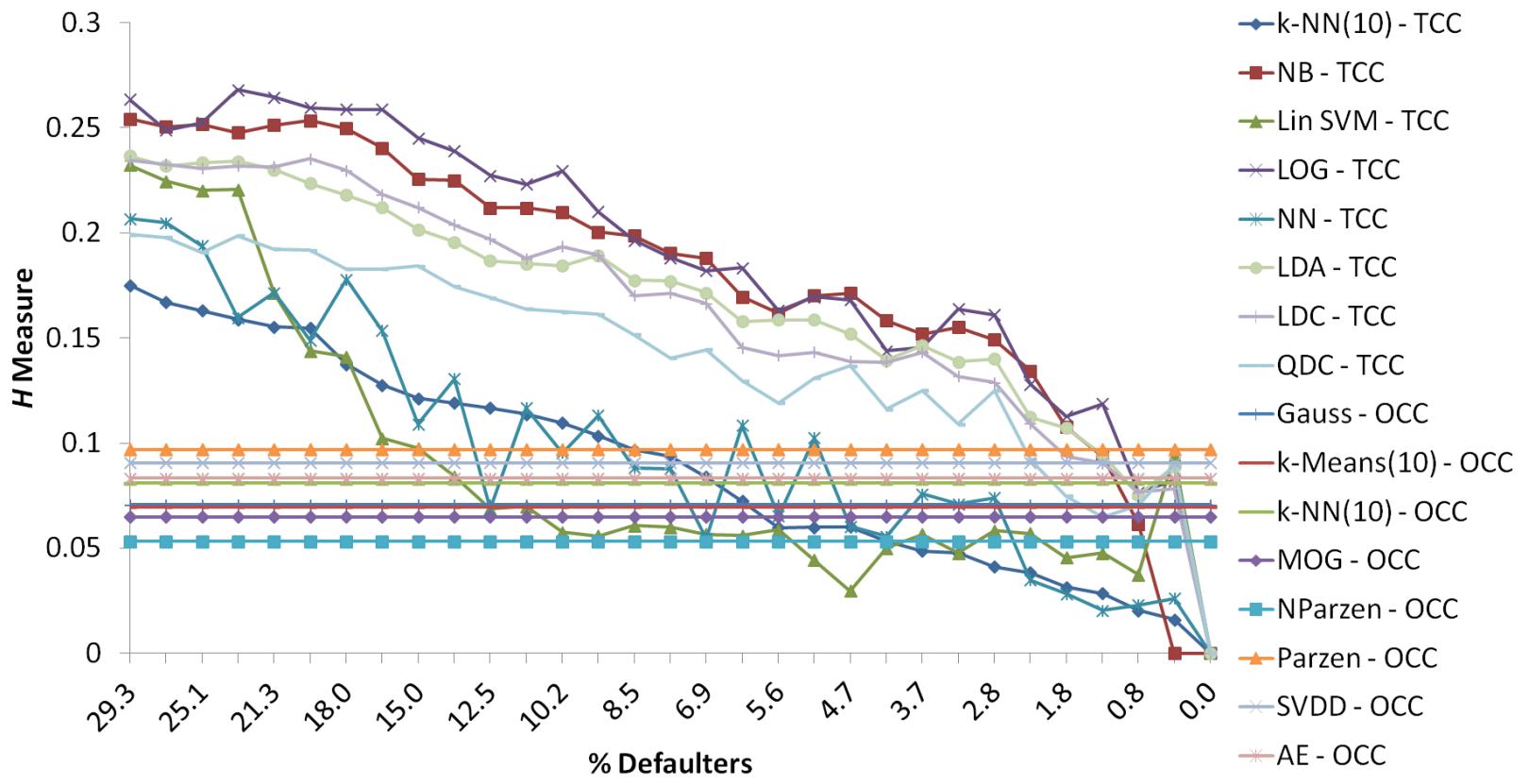


Figure 5.4: German: Normal process and one-class classification process test set H measure performance.

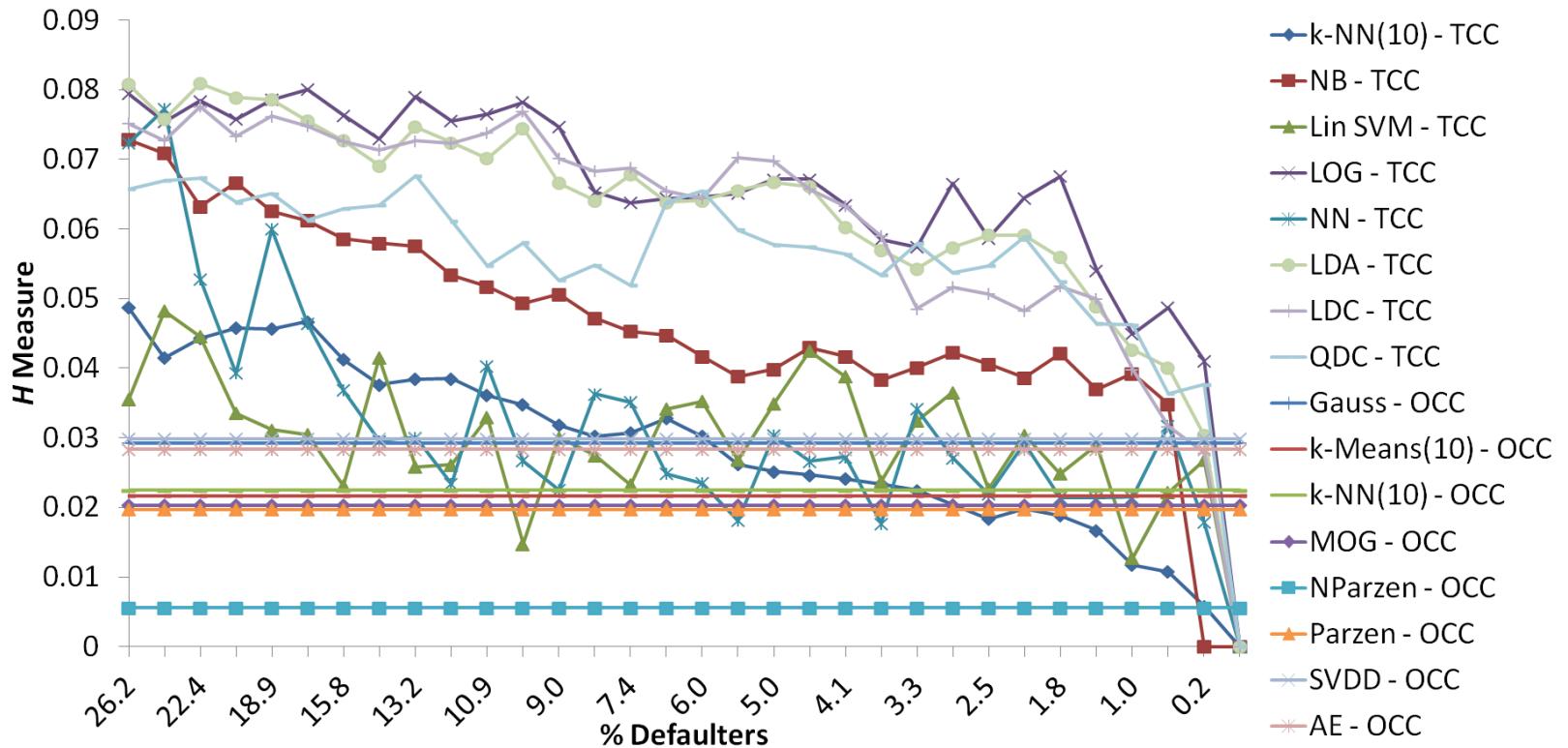


Figure 5.5: Thomas: Normal process and one-class classification process test set H measure performance.

5.2.1 Two-class Classifier Performance with Imbalance

The effects of class imbalance using the normal process are clearly evident in Figure 5.3. Beginning with a class imbalance of 44:56 (44% non-defaulter, 56% defaulter) the performance of the two-class classifiers gradually deteriorates as the class imbalance increases through the removal of instances from the defaulter class in the training set. The performance of the naïve Bayes and logistic regression classifiers remains relatively robust while, in contrast, the performance of the Lin SVM, k -NN and NN classifiers deteriorates rather more rapidly. So that more general comparisons can be made, Table 5.2 shows the H measure for each two-class classifier at imbalance ratios of 70:30, 80:20, and 90:10 for the nine datasets used. The H measure values vary by dataset, with particularly poor performance recorded for the PAKDD, Thomas, and Spain datasets. Poor classifier performance, as measured by the AUC, has previously been reported for the PAKDD (see Xie *et al.* (2009)) and Thomas (see Wang *et al.* (2005)) datasets. Reasons for this relatively poor performance are explained later in the text.

Table 5.2 confirms that the performance of naïve Bayes, logistic regression, LDA, LDC and, to a slightly lesser extent, QDC remain, for the most part, robust even as far as a class imbalance of 90:10. In comparison, at 90:10, the performance of the NN, Lin SVM and k -NN begin to languish. For each dataset and class imbalance ratio in Table 5.2 we compute a ranking of the different classifiers assigning rank 1 to the classifier yielding the best test set H measure, and rank 8 to the classifier giving the worst test set H measure. The average ranking of each classifier over the three selected class imbalance ratios is computed. This figure is then averaged over

the nine datasets and reported as the *average rank*. Based on average rank logistic regression performs best. This should come as no surprise as both (Baesens *et al.*, 2003) and (Xiao *et al.*, 2006) reported logistic regression as performing strongly when assessed for credit scoring problems.

At 70:30, the differences in performance of logistic regression and the other supervised classifiers were compared using a Friedman test. No statistically significant differences between the groups were observed. This is not surprising as previous studies (Baesens *et al.*, 2003) have reported that the majority of classification techniques yield classification performances that are quite competitive with each other. At 80:20 *k*-NN, NN and QDC perform significantly worse than logistic regression. At 90:10, the performance of Lin SVM shows a marked deterioration compared to QDC, resulting in *k*-NN, NN and Lin SVM performing significantly worse than logistic regression. It should be noted that the performance range of the *H* measure values varies considerably from dataset to dataset. Normally, *H* measure values range from zero for models which randomly assign class labels, to one for models which obtain perfect classification. Such variation suggests that some of the datasets used in the evaluation may be less discriminable than others. Unlike the AUC or Gini coefficient, the *H* measure does depend on the class priors. Hence, for two given datasets the *H* measure may be different because of two effects¹. Firstly, the classification performance may be different, due to the discriminating power of the attributes. Secondly, the class distribution can be different, affecting the *H* measure through the class priors. As the datasets display different degrees of skewness, it seems likely that the difference in *H* measure values is caused by a mixture of both

¹A comment from an anonymous reviewer

effects.

To summarise, our findings show that the performance of two-class classifiers deteriorates as class imbalance increases - highlighting the reason that LDPs are such a problem. Up as far as a class imbalance ratio of 90:10 the rate of deterioration in the performance of many of the two-class classifiers is gradual with no sudden decreases. Some of the classifiers (naïve Bayes, logistic regression, LDA, LDC) remain relatively robust to class imbalance. Ultimately, however, at a very high level of class imbalance the two-class classifiers succumb to a poor classification performance.

Table 5.2: Test set H measure performance using the normal process on two-class classifiers. The best test set H measure at each class imbalance ratio is underlined. The rank of the different classifiers at each class imbalance ratio is averaged over all the datasets and reported as the AR (average rank). For legibility the H measure figures have been scaled and should be multiplied by 10^{-2} .

Technique	Australia			German			Poland			UCSD			Iran		Japan		PAKDD		Thomas		Spain	AR Total
	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10
k -NN(10)	64.1	59.9	50.6	17.5	15.5	11.0	29.3	24.7	19.0	36.0	32.8	27.1	28.2	21.7	11.1	12.1	2.0	1.0	4.6	3.5	2.0	7.2
NB	64.0	62.8	59.2	25.4	25.3	21.0	37.4	35.3	31.0	45.8	45.4	43.6	47.0	44.0	<u>41.6</u>	<u>35.7</u>	4.8	4.2	6.7	4.9	5.1	3.7
Lin SVM	70.4	70.0	61.4	23.2	14.4	5.8	39.5	38.6	<u>37.0</u>	47.1	46.9	40.8	46.4	43.4	39.2	25.2	0.9	0.7	3.4	1.5	1.1	4.9
LOG	65.8	63.9	63.5	<u>26.3</u>	<u>26.0</u>	<u>23.0</u>	38.0	<u>38.8</u>	36.5	49.2	<u>47.9</u>	<u>44.8</u>	<u>54.0</u>	<u>52.9</u>	34.2	24.0	5.7	5.2	7.6	<u>7.8</u>	4.9	<u>2.3</u>
NN	68.4	61.7	50.6	20.7	14.9	9.5	<u>40.9</u>	37.3	35.7	<u>49.5</u>	45.4	33.7	41.2	40.6	22.7	18.2	3.5	2.3	3.9	2.7	2.0	5.7
LDA	70.9	<u>70.1</u>	<u>69.3</u>	23.7	22.4	18.4	27.2	27.9	22.2	46.4	45.9	43.7	43.4	40.5	31.6	28.5	6.0	5.1	<u>7.9</u>	7.4	<u>5.2</u>	3.4
LDC	<u>71.6</u>	69.1	66.9	23.4	23.5	19.4	25.8	24.9	22.1	46.2	45.7	43.3	46.7	44.0	32.3	21.8	<u>6.1</u>	<u>5.6</u>	7.3	7.7	4.9	3.7
QDC	70.7	68.3	64.8	19.9	19.2	16.2	35.1	32.7	29.1	45.0	44.9	43.4	34.4	33.0	30.6	26.0	3.7	4.1	6.4	5.8	4.4	5.1

Table 5.3: Test set H measure performance using the oversample process on two-class classifiers. The best test set H measure for each class imbalance ratio is underlined. The rank of the different classifiers at each class imbalance ratio is averaged over all the datasets and reported as the AR (average rank). For legibility the H measure figures have been scaled and should be multiplied by 10^{-2} .

Technique	Australia			German			Poland			UCSD			Iran		Japan		PAKDD		Thomas		Spain	AR Total
	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	
k -NN(10)	62.3	54.3	48.5	18.1	13.5	13.2	32.4	27.8	24.2	35.0	31.6	26.2	29.7	22.4	12.3	15.9	2.4	1.1	4.7	4.3	1.4	7.5
NB	61.3	56.3	50.6	25.3	22.1	17.3	38.3	34.7	21.9	45.8	45.0	41.2	45.8	37.8	<u>36.4</u>	<u>29.9</u>	5.6	4.7	7.1	6.6	<u>5.3</u>	4.3
Lin SVM	69.7	62.5	64.0	23.8	21.7	17.8	38.1	<u>39.5</u>	35.9	46.9	46.7	44.8	44.6	41.2	25.5	13.9	5.9	5.2	6.8	7.8	4.7	3.7
LOG	63.1	60.3	56.7	<u>25.5</u>	<u>24.1</u>	<u>21.3</u>	<u>39.7</u>	38.3	<u>36.7</u>	<u>50.1</u>	<u>48.8</u>	<u>45.1</u>	<u>54.5</u>	<u>51.0</u>	33.8	26.7	5.3	4.8	<u>7.9</u>	7.8	5.0	<u>2.4</u>
NN	66.9	61.4	59.9	20.3	18.4	14.5	32.2	34.2	32.5	49.5	47.0	43.2	40.5	33.7	27.4	17.0	4.4	4.0	6.5	6.1	3.4	5.3
LDA	<u>71.0</u>	<u>65.7</u>	<u>67.8</u>	23.6	21.7	18.9	25.3	25.4	19.8	46.2	45.7	43.1	41.3	38.1	32.0	21.3	<u>6.1</u>	<u>5.6</u>	7.7	7.9	5.0	3.6
LDC	<u>71.0</u>	<u>65.7</u>	<u>67.8</u>	23.6	21.7	18.9	25.6	28.3	22.5	46.2	45.7	43.3	43.9	41.1	32.0	21.3	6.0	<u>5.6</u>	7.6	<u>8.0</u>	5.0	3.4
QDC	70.3	62.9	66.0	20.7	17.4	16.6	27.1	30.8	29.6	44.9	45.0	43.3	32.5	30.1	25.9	26.3	4.0	3.5	6.6	6.2	4.5	5.4

5.2.2 The Impact of Oversampling

Figure 5.6, Figure 5.7, and Figure 5.8 illustrate the resulting H measure when the eight two-class classifiers using the oversample process were tested on the Australia, German, and Thomas datasets, respectively. Similarly, for the remaining datasets, H measure performance of classifiers using the oversample process is displayed in Figure C.7 to Figure C.12 in Appendix C. Figures C.39 to C.47 in Appendix C are included to demonstrate the performance of classifiers measured using the AUC under the oversample process.

With reference to the aforementioned figures (Figure 5.6, Figure 5.7, and Figure 5.8), when contrasted with the normal process, the oversampling process appears to improve the performance of the weaker two-class classifiers. For example, comparing the results from the Thomas dataset in Figure 5.5 (normal process) with Figure 5.8 (oversample process), Lin SVM and NN improve substantially. In contrast, the better performing two-class classifiers from the normal process (e.g. logistic regression) do not appear to benefit from the oversampling process.

The results of the oversampling process are detailed in Table 5.3. As previously reported, oversampling improves the performance of the weaker two-class classifiers - NN, Lin SVM, and k -NN - but fails to raise the performance of the stronger ones - naïve Bayes, logistic regression, LDA, LDC and QDC. In fact naïve Bayes, QDC and LDA show a decline in performance. Kolcz *et al.* (2003) previously reported that at high levels of data duplication the performance of naïve Bayes deteriorates. As per the normal process, logistic regression performs best based on the average rank. At 70:30 no statistically significant difference between the classifiers is detected. At

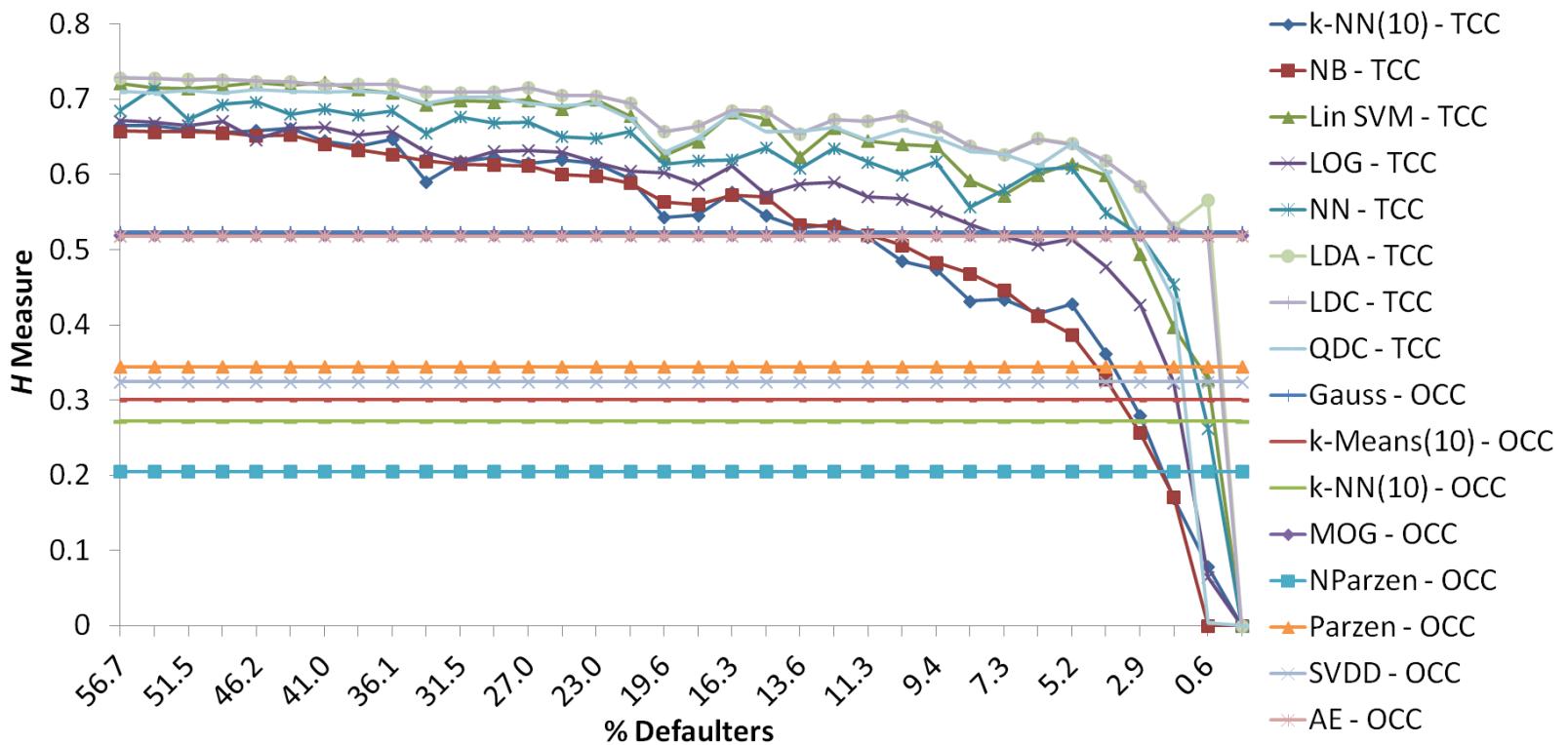


Figure 5.6: Australia: Oversample process and one-class classification process test set H measure performance.

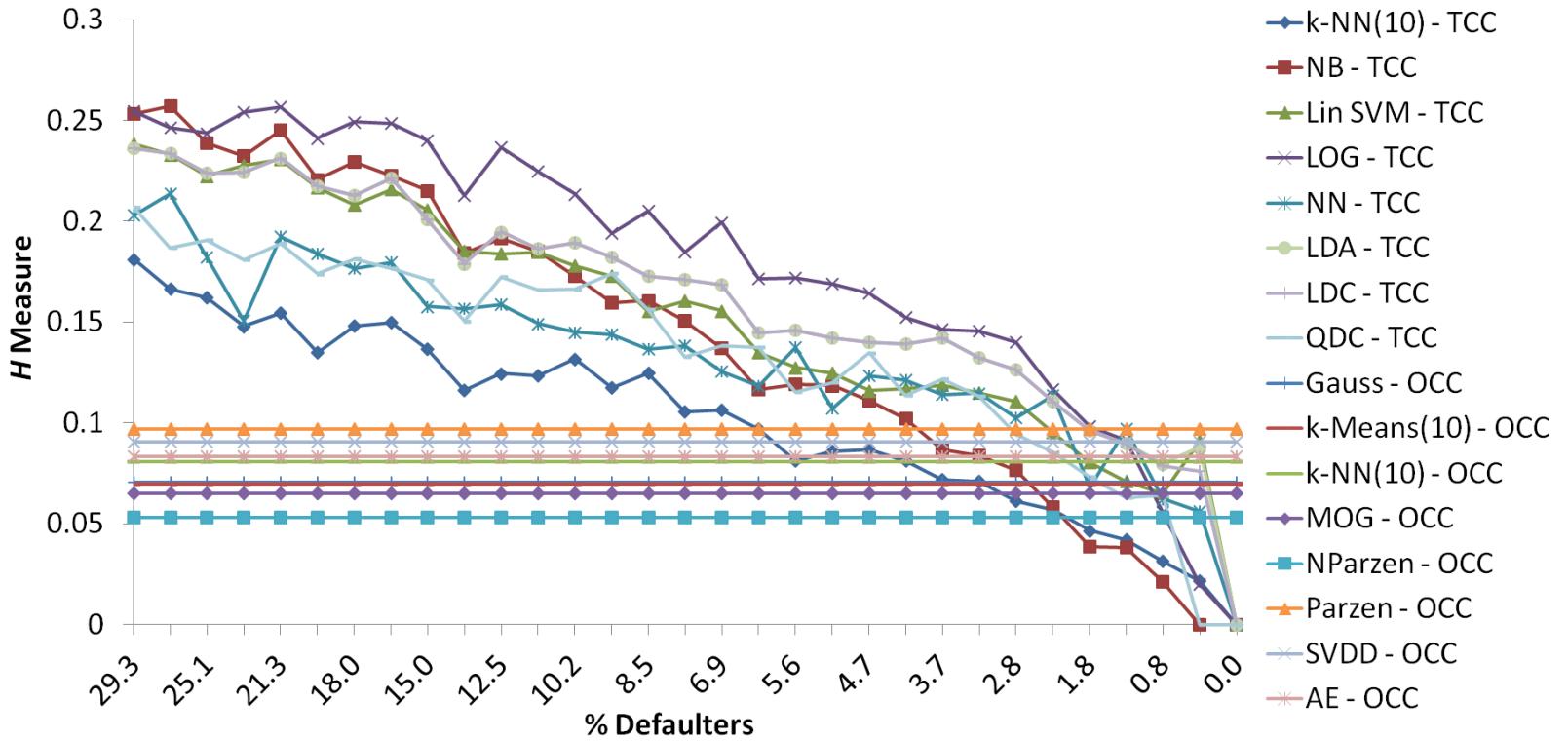


Figure 5.7: German: Oversample process and one-class classification process test set H measure performance.

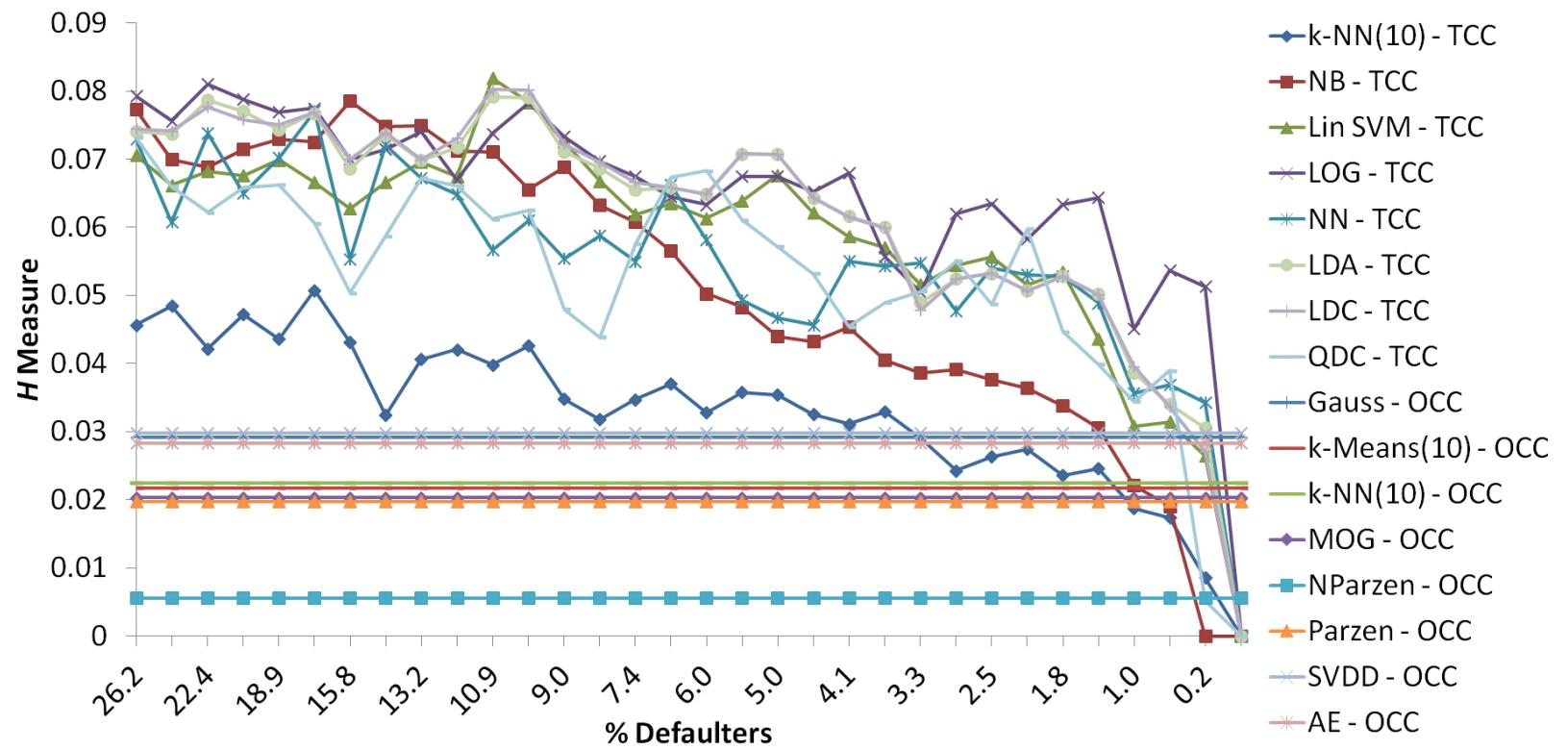


Figure 5.8: Thomas: Oversample process and one-class classification process test set H measure performance.

80:20, k -NN and QDC perform significantly worse than logistic regression. At 90:10,

k -NN and NN perform significantly worse than logistic regression.

Table 5.4 compares the difference between the normal process and oversample process averaged over the nine datasets at each of the three separate class imbalance ratios. A positive figure indicates that the oversample process performed better than the normal process. Lin SVM shows the largest improvement, this is generated in part by the oversample process performance on the PAKDD and Thomas datasets. We conjecture that the reason for this large improvement in performance is the increase in the number of support vectors and the fixed cost parameter. Even though the performance of NN and k -NN improve with oversampling, it is insufficient to make a statistically significant difference. The best performing two-class classifier, logistic regression, shows a small decline in performance when the data is duplicated.

Table 5.4: Average difference in test set H measure performance, oversample process *versus* normal process. A positive figure indicates the oversample process outperformed the normal process.

Technique	70:30	80:20	90:10
k -NN(10)	2%	4%	8%
NB	-1%	-2%	-5%
Lin SVM	-1%	83%	171%
LOG	0%	-2%	-2%
NN	-6%	16%	38%
LDA	-2%	-3%	-3%
LDC	0%	0%	0%
QDC	-5%	-4%	-1%

To summarise, our findings show that oversampling improves the performance of the two-class classifiers worst affected by class imbalance. However, the performance of the more robust two-class classifiers displays no overall benefit from oversampling, suggesting that it is not an appropriate solution to the LDP problem.

5.2.3 One-class Classifiers

The next step of the evaluation is to compare the classification performance of the two-class classifiers with that of the one-class classifiers. Based on Figure 5.3, the cross-over in performance between the best two-class classifier and best one-class classifier occurs at a high level of imbalance, typically 99:1 (i.e. 99% non-defaulter, 1% defaulter). We select this class imbalance ratio in order to best mirror the LDP problem.

Based on the results presented in Sections 5.2.1 and 5.2.2 logistic regression using the normal process (LOG_Norm) is the classifier that is taken forward for comparison with a selection of one-class classifiers at a class imbalance ratio of 99:1. The level of class imbalance does not affect the one-class classifiers as they do not employ non-target data during training.

Table 5.5 reports the H measure performance for LOG_Norm at a class imbalance of 99:1, along with the one-class classifiers using the OCC process. The challenging nature of the LDP problem is underscored by the low H measure scores reported for the German, PAKDD, Spain, and Thomas datasets. The average ranking of the classifiers over the nine datasets is also provided which shows that LOG_Norm performs best.

Even at such a high imbalance of 99:1, LOG_Norm performs competitively with the one-class classifiers. The OCC process outperforms LOG_Norm on 5 of the 9 datasets albeit with different OCC classifiers.

To summarise, no evidence exists from our experimentation to show that one-class classification outperforms two-class classification with differences that are sta-

Table 5.5: Test set H measure performance of logistic regression normal process (LOG_Norm), and OCC process at a class imbalance ratio of 99:1. The best test set H measure per dataset is underlined. The average rank (AR) of the classifiers is also provided. H measure figures should be multiplied by 10^{-2} . *Aus* = Australia, *Ger* = German.

Technique	<i>Aus</i>	<i>Ger</i>	<i>Iran</i>	<i>Japan</i>	<i>PAKDD</i>	<i>Poland</i>	<i>Spain</i>	<i>Thomas</i>	<i>UCSD</i>	<i>AR</i>
LOG_Norm	50.1	7.7	<u>30.5</u>	20.5	1.8	<u>26.9</u>	<u>2.3</u>	<u>4.5</u>	40.1	<u>2.8</u>
Gauss	<u>52.3</u>	7.1	4.6	25.8	1.6	15.7	1.1	2.9	35.3	3.3
<i>k</i> -Means(10)	30.0	7.0	5.5	28.0	1.3	8.1	1.0	2.2	21.2	5.4
<i>k</i> -NN(10)	27.2	8.1	3.9	23.3	0.9	4.7	0.8	2.2	23.3	6.6
MoG	51.9	6.5	2.8	19.7	1.4	7.6	0.9	2.0	40.6	6.3
NParzen	20.5	5.3	5.8	19.8	0.3	14.9	1.0	0.6	<u>40.8</u>	6.1
Parzen	34.4	<u>9.7</u>	3.6	25.6	1.2	8.2	0.6	2.0	25.8	5.9
SVDD	32.5	9.1	5.4	23.1	1.8	14.5	1.0	3.0	22.6	4.2
AE	51.8	8.3	3.7	<u>31.1</u>	<u>2.2</u>	10.4	0.9	2.8	17.7	4.3

tistically significant. In some ways this is to be expected as the two-class classifiers use more instances during training. However, the fact that OCC outperforms two-class classifiers on a majority of our selected datasets indicates that, under an extreme imbalance (a defaulter class rate of 1% or lower) one should consider employing OCC as an approach to addressing the LDP problem.

5.2.4 Optimising the Threshold

In this section we illustrate the appropriateness of adjusting the threshold on classifier output as a means of addressing the class imbalance problem. In practice it is necessary to select a threshold on classification output in order to make actual classifications. The validation dataset is used to identify an optimised threshold for both the one- and two-class classifiers. When a classification threshold is used we use the harmonic mean to measure performance.

Table 5.6 compares the performance of the two-class classifiers using a standard threshold of 0.5 and an optimised threshold on two datasets (Australia and German)

at a class imbalance of 90:10. The Australia and German datasets are selected as they are the most commonly used datasets in credit risk scoring literature. Based on the performance of k -NN(10), Lin SVM, and LDA, using the standard threshold of 0.5, for these classifiers, is clearly inappropriate for the German dataset.

Table 5.6: Test set harmonic mean performance of Default threshold (D) *versus* Optimised threshold (O) at a class imbalance ratio 90:10 using the Australia (*Aus*) and German (*Ger*) datasets. Harmonic mean figures should be multiplied by 10^{-2} .

<i>Technique</i>	<i>Aus (D)</i>	<i>Aus (O)</i>	<i>Ger (D)</i>	<i>Ger (O)</i>
k -NN(10)	39.9	80.7	2.0	58.3
NB	83.6	83.5	67.5	67.5
Lin SVM	45.5	78.8	0.0	50.2
LOG	85.2	85.2	23.6	67.2
NN	68.4	79.1	34.9	57.4
LDA	69.2	87.6	0.9	65.5
LDC	86.7	86.9	66.8	67.8
QDC	82.9	84.5	62.7	63.3

In all but three of the constructed models, the optimised threshold improves the performance of the two-class classifiers. This supports the recommendation of previous studies (Provost, 2000; Vinciotti & Hand, 2003) which cite that adjusting the threshold is the most straight-forward approach to dealing with imbalanced datasets. Based on the harmonic mean performance measure, we compare the performance of two-class classifiers using the normal process when an optimised threshold is used. Figure 5.9 illustrates the classification performance on the Australia dataset, measured using the harmonic mean, of the two-class classifiers using the normal process and the one-class classifiers using the OCC process. Similarly, Figures C.13 to C.20 in Appendix C are included for the remaining datasets. When compared with the H measure performance, the harmonic mean performance of many of the classifiers using an optimised threshold only begins to taper away at high level of class

imbalance.

Table 5.7 displays the results of the two-class classifiers using the normal process across our three selected class imbalances of 70:30, 80:20 and 90:10. The harmonic mean performance of the classifiers using an optimised threshold remains stable across the selected class imbalances. The deterioration in performance arising from class imbalance is not as immediate as when a default threshold is used.

As per Table 5.2, the average ranking of the two-class classifiers across the selected class imbalances is provided in Table 5.7. When the classifiers' average ranking based on the H measure (Table 5.2) is compared to that of the harmonic mean (Table 5.7), five of the eight classifiers attain the same average ranking. This indicates a satisfactory degree of consistency between the performance measures. Logistic regression performs best, as per the previous experiments. The performance of the two-class classifiers declines as the class imbalance increases. At a class imbalance ratio of 70:30 no statistical significance between the two-class classifiers is detected. At 80:20, significance is detected, with logistic regression outperforming NN. At 90:10 the performance of k -NN, Lin SVM and NN are inferior to that of logistic regression.

Table 5.8 displays the results for the oversample process. The average ranking of the oversampled two-class classifiers reveals that logistic regression performs best again. At a class imbalance of 70:30 no statistically significant difference is detected between the oversampled classifiers but at the 80:20 and 90:10 class imbalance ratios, k -NN and QDC perform significantly worse than logistic regression. As per the normal process, when the classifiers' average ranking based on the H measure (Table 5.3) is compared to that of the harmonic mean (Table 5.8), five of the eight classifiers

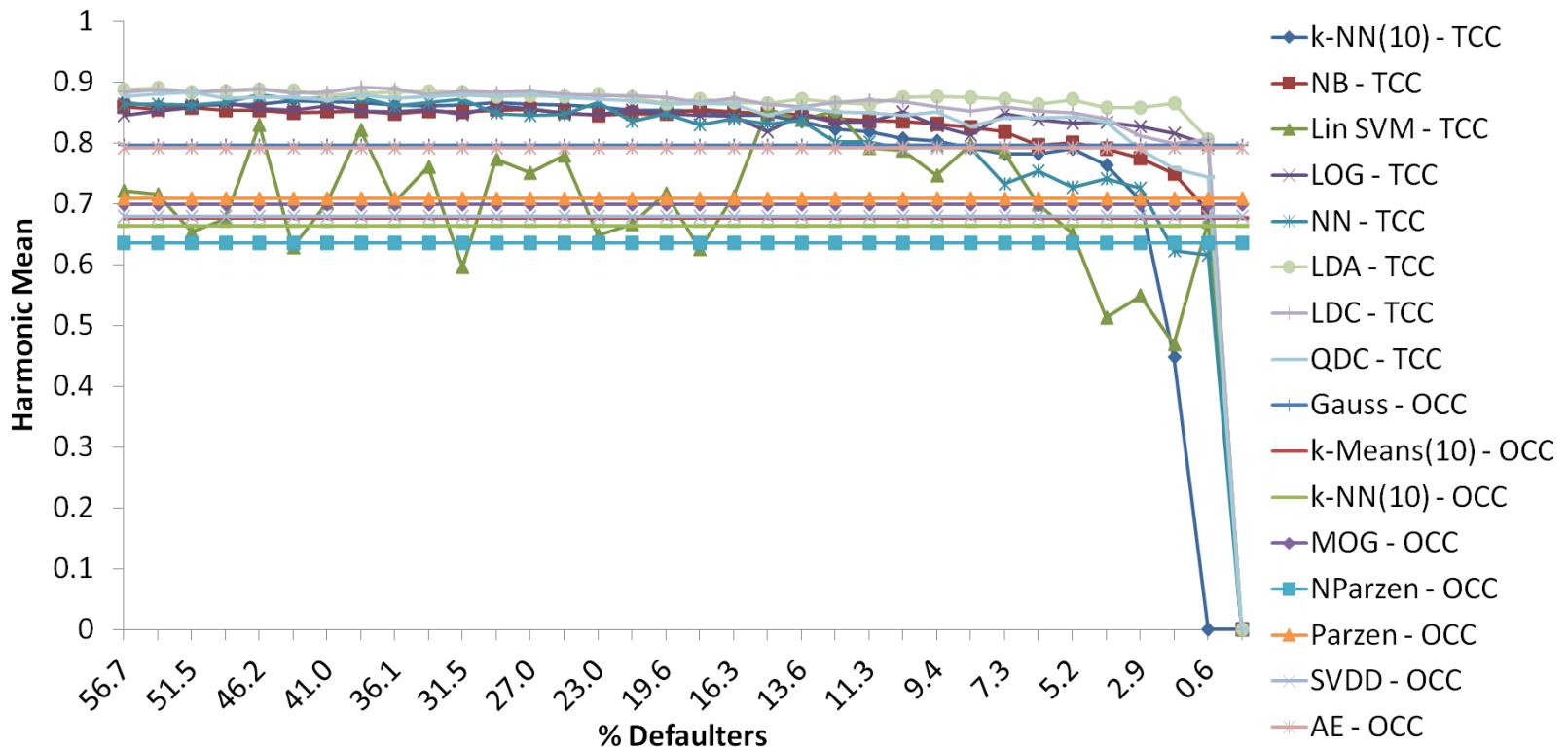


Figure 5.9: Australia: Normal process and one-class classification process test set harmonic mean performance.

attain the same average ranking. Figures C.21 to C.29 in the Appendix C are included to demonstrate the harmonic mean performance of the two-class classifiers using the oversample process and the one-class classifiers using the OCC process.

Table 5.7: Test set harmonic mean performance using the normal process on two-class classifiers. The best test set harmonic mean for each class imbalance ratio is underlined. The rank of the different classifiers at each class imbalance ratio is averaged over all the datasets and reported as the AR (average rank). For legibility the harmonic mean figures have been scaled and should be multiplied by 10^{-2} .

Technique	Australia			German			Poland			UCSD			Iran		Japan		PAKDD		Thomas		Spain	AR Total
	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10
k-NN(10)	86.7	85.4	80.7	66.0	63.5	58.3	68.9	<u>66.8</u>	62.4	75.1	73.6	70.5	72.6	71.4	48.5	54.2	52.7	45.6	52.9	51.2	52.9	5.9
NB	85.5	84.8	83.5	69.3	<u>70.0</u>	67.5	67.3	66.2	65.9	78.7	78.2	77.7	75.3	74.5	<u>67.2</u>	<u>63.1</u>	56.9	55.9	<u>58.5</u>	<u>58.6</u>	64.2	3.6
Lin SVM	77.3	71.7	78.8	68.7	30.4	50.2	67.0	65.1	63.5	79.8	79.7	77.1	77.1	74.5	58.5	50.8	17.7	40.0	40.6	42.8	44.1	6.1
LOG	86.1	84.9	85.2	69.0	69.2	67.2	68.7	65.3	64.0	80.7	<u>80.1</u>	<u>78.9</u>	<u>80.9</u>	<u>79.6</u>	62.2	61.6	57.3	56.1	58.4	57.7	62.6	<u>2.9</u>
NN	84.9	84.8	79.1	66.4	60.0	57.4	<u>70.1</u>	65.6	<u>67.6</u>	<u>80.7</u>	79.2	73.7	74.1	<u>75.7</u>	45.8	41.1	52.0	52.5	52.5	49.0	52.8	5.6
LDA	87.7	86.6	<u>87.6</u>	<u>69.9</u>	68.7	65.5	65.3	60.8	61.4	79.8	79.5	78.1	76.5	74.4	63.2	58.4	59.0	<u>59.0</u>	58.2	58.1	<u>64.4</u>	3.3
LDC	<u>88.4</u>	<u>87.5</u>	86.9	69.5	68.1	<u>67.8</u>	61.4	64.4	60.2	79.6	79.5	78.2	76.6	73.1	64.2	52.5	<u>59.5</u>	58.5	58.1	58.2	64.3	3.3
QDC	87.7	86.4	84.5	66.9	66.7	63.3	63.1	62.4	56.7	78.8	78.7	78.2	74.5	68.7	59.2	51.8	56.5	54.7	55.4	54.2	61.8	5.3

Table 5.8: Test set harmonic mean performance using the oversample process on two-class classifiers. The best test set harmonic mean for each class imbalance ratio is underlined. The rank of the different classifiers at each class imbalance ratio is averaged over all the datasets and reported as the AR (average rank). For legibility the harmonic mean figures have been scaled and should be multiplied by 10^{-2} .

Technique	Australia			German			Poland			UCSD			Iran		Japan		PAKDD		Thomas		Spain	AR Total
	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	70:30	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10	80:20	90:10
k-NN(10)	84.6	81.1	80.2	63.9	61.6	62.1	<u>70.0</u>	63.1	62.4	74.2	72.6	63.9	72.1	66.0	52.0	52.6	53.8	52.4	52.3	55.0	55.0	7.1
NB	83.9	82.1	80.3	69.7	67.2	64.1	68.7	64.2	<u>57.9</u>	78.8	78.5	77.3	75.2	71.3	59.4	<u>59.4</u>	<u>58.2</u>	57.1	56.1	55.3	63.6	4.7
Lin SVM	79.7	84.1	85.9	70.1	67.0	64.8	66.7	<u>69.0</u>	<u>67.1</u>	79.6	79.3	78.9	74.7	74.0	59.2	46.5	54.2	<u>58.3</u>	58.0	58.0	63.7	3.8
LOG	83.7	81.7	81.1	<u>70.3</u>	<u>69.5</u>	<u>67.2</u>	67.1	66.0	65.7	<u>81.2</u>	<u>80.5</u>	<u>79.3</u>	<u>80.3</u>	<u>77.9</u>	<u>61.6</u>	58.1	57.9	55.1	58.4	57.5	<u>64.7</u>	<u>2.6</u>
NN	87.2	83.2	82.2	66.5	65.5	62.6	66.2	65.3	64.9	80.7	80.1	78.4	71.8	72.6	57.3	43.6	56.1	56.0	57.7	57.4	58.6	5.2
LDA	87.6	<u>85.8</u>	<u>86.4</u>	69.9	67.0	67.1	61.6	63.3	56.8	79.7	79.1	78.3	75.1	72.4	60.8	53.3	57.4	58.1	59.4	58.2	63.5	3.5
LDC	87.6	<u>85.8</u>	<u>86.4</u>	69.9	67.0	67.1	62.6	65.8	60.0	79.7	79.2	78.6	74.7	73.0	60.8	53.3	57.3	58.1	<u>59.5</u>	<u>58.3</u>	63.5	3.0
QDC	<u>88.0</u>	83.9	86.4	66.8	65.7	65.2	60.2	63.7	59.4	78.6	78.5	78.0	73.2	69.4	54.7	51.1	57.2	54.8	54.9	54.5	61.5	5.8

The average difference, over the nine datasets, between the normal process and oversample process of the two-class classifiers is displayed in Table 5.9. Based on the harmonic mean at the class imbalance ratio of 70:30 oversampling makes no overall difference to the performance of the two-class classifiers. At higher levels of class imbalance, the non-parametric classifiers NN and Lin SVM benefit from oversampling, as observed previously. Again, the performance of naïve Bayes is somewhat impeded by oversampling and the best performance for logistic regression occurs using the normal process rather than the oversample process.

Table 5.9: Average difference in test set harmonic mean performance, oversample process versus normal process. Positive figure indicates oversample process outperformed normal process.

<i>Technique</i>	<i>70:30</i>	<i>80:20</i>	<i>90:10</i>
<i>k</i> -NN(10)	-1%	-1%	1%
NB	0%	-3%	-4%
Lin SVM	1%	49%	18%
LOG	-1%	0%	-1%
NN	-1%	6%	6%
LDA	-1%	-1%	-2%
LDC	0%	-1%	0%
QDC	-1%	-1%	1%

We next compare logistic regression using the normal process to a selection of one-class classifiers at an imbalance of 99:1. Optimised thresholds are calculated for all techniques used. The results of this comparison are displayed in Table 5.10 and are, in general, very similar to the results in Section 5.2.3. Logistic regression remains the best performing classifier. Further to the results of the H measure, in which logistic regression performs significantly better than *k*-NN and MOG, the performance of logistic regression with an optimised classification threshold is significantly better than a number of one-class classifiers (including SVDD, naïve Parzen,

mixture of Gaussians and k -NN). Of the one-class classifiers, based on average ranking, the Gaussian performs best. Based on the average ranking, it is worth noting that the harmonic mean performance of the SVDD classifier is somewhat worse compared to its corresponding H measure performance. This difference highlights the sensitivity of selecting appropriate SVDD parameters at a specific classification threshold.

Table 5.10: Test set harmonic mean performance of logistic regression normal process (LOG_Norm), and OCC process at a class imbalance ratio of 99:1. The best test set harmonic mean per dataset is underlined. The average rank (AR) of the classifiers is also provided. Harmonic mean figures should be multiplied by 10^{-2} . *Aus* = Australia, *Ger* = German.

<i>Technique</i>	<i>Aus</i>	<i>Ger</i>	<i>Iran</i>	<i>Japan</i>	<i>PAKDD</i>	<i>Poland</i>	<i>Spain</i>	<i>Thomas</i>	<i>UCSD</i>	<i>AR</i>
LOG_Norm	<u>79.8</u>	58.5	<u>73.5</u>	53.3	52.4	<u>64.4</u>	<u>57.1</u>	<u>55.4</u>	76.4	<u>2.1</u>
Gauss	79.6	55.9	51.5	56.8	53.3	57.9	52.3	52.4	73.9	3.0
k -Means(10)	67.7	56.4	54.8	57.1	52.0	47.2	51.6	49.9	67.0	5.1
k -NN(10)	66.4	55.6	50.0	<u>58.6</u>	51.1	42.1	51.9	48.5	68.9	5.7
MoG	69.9	55.2	46.1	49.6	39.7	46.9	45.1	46.6	73.0	7.4
NParzen	63.5	53.8	46.8	49.8	46.3	56.0	53.3	44.4	<u>77.4</u>	6.1
Parzen	70.9	<u>58.8</u>	49.2	55.2	50.4	48.6	52.5	48.1	69.7	4.9
SVDD	68.0	57.5	34.3	56.3	52.2	50.7	50.7	51.5	68.3	5.7
AE	79.3	55.4	44.3	56.8	<u>55.4</u>	53.9	51.1	53.9	65.5	5.0

To summarise, selecting an appropriate threshold can substantially improve the performance of a two-class classifier. However, it is worth noting that this cut-off decision is dependent on a number of factors, including: (i) what aspect of classifier performance is being examined; (ii) the relative cost ratio of false positives and false negatives; and (iii) the strategic considerations of the bank (e.g. how to pool loans into different risk grades). For these reasons it is common for financial institutions to use the ROC performance measure to assess classifier performance over a range of cut-offs. Further discussion of the results with respect to other empirical studies is reserved in the Conclusions of this chapter.

5.3 Conclusions

This chapter presented an extensive evaluation of approaches to addressing the LDP problem when building credit scoring models. Based on our findings presented in this chapter, we believe that when both target and non-target data is available, the semi-supervised OCC techniques should not be expected to outperform the supervised two-class classification techniques. This is based on the fact that two-class classifiers use more information during training.

Even though it cannot be unanimously proven that OCC is better than two-class classification at very low levels of defaulters, the performance of OCC merits consideration as a solution to the LDP problem. Based on the H measure performance measure, OCC outperforms logistic regression on 5 of the 9 datasets. Similarly, based on the harmonic mean performance measure, OCC outperforms logistic regression on 4 of the 9 datasets. This indicates that, under an extreme imbalance (a defaulter class rate of 1% or lower) one should consider employing OCC as an approach to addressing the LDP problem.

Sampling is one of the simplest and most popular solutions to the class imbalance problem. Although oversampling improves the performance of some two-class classifiers, it does not lead to an overall improvement of the best performing classifiers - that is the strong do not become stronger. In fact, in our experiment the performance of the best performing two-class classifier, logistic regression, registered a small decline when oversampling was applied, which matches the results of Bellotti & Crook (2008) and Crone & Finlay (2012) with the latter reporting that oversampling “*appears to be of minor importance*” with respect to the performance

of logistic regression.

Based on these findings, oversampling should not be employed with logistic regression as a suitable technique to address the LDP problem. As oversampling does not introduce any new data, the fundamental “*lack of data*” issue is not addressed (Burez & Van den Poel, 2009).

Adjusting the threshold on classification output yields a large improvement in classifier performance. It is therefore advisable, in addressing the low-default portfolio problem, to optimise the classification threshold before pursuing some of the more sophisticated methods associated with data sampling and cost sensitive learning.

Although many studies discuss the importance of classification threshold selection (see Baesens *et al.*, 2003), very few actually conduct any sort of assessment of the predictive performance of classifiers using an optimised classification threshold. Many studies sidestep the problem of choosing a specific classification threshold by using the AUC. The comprehensiveness of this study is enhanced by employing a threshold specific performance measure such as the harmonic mean.

Our findings also match Lee & Cho (2007) who performed a modest comparison of one- and two- class classifiers for response modelling and found that with a response rate (the minority class) of 1% or lower one should apply OCC to the majority class. However, with respect to oversampling, our findings are contradicted, somewhat, by Marqués *et al.* (2012) who report that random oversampling improves classifier performance when using imbalanced data. The results presented in this chapter were based on experiments performed using 9 real-world datasets and 8 two-class classifiers. In comparison, Marqués *et al.* (2012) used 5 real-world datasets (one of which should be discounted as it is the incorrect version of the Japan dataset),

and 2 two-class classifiers (logistic regression and Lin SVM). Marqués *et al.* (2012)

do not examine the effects of oversampling below a defaulter class rate of 5%. For researchers interested in LDPs it would be interesting to see the effects of oversampling assessed at lower defaulter class rates as, typically, LDPs contain very few or no defaulters. It is worth noting that for the German dataset, the largest AUC value reported by Marqués *et al.* (2012) for the linear kernel SVM classifier is 0.51 which is recorded based on a 20% default rate in the dataset. For higher class imbalances (default rates of 14.3% to 6.67%) this figure is 0.5. In contrast, Brown & Mues (2012) report an AUC value of 0.768 for the German dataset when a default rate of 10% is used. This may indicate that the parameters of the linear kernel SVM classifier used by Marqués *et al.* (2012) were inadequately optimised. Overall, such discrepancies reflect the fact that it is not uncommon for studies to differ on the question of oversampling (see Chawla *et al.*, 2004; Drummond & Holte, 2004), and further research is required as there is no single final answer to the question. Our work, however, has contributed concrete and meaningful results to the specific context of low-default portfolios, and also to the topic of class imbalance in credit scoring.

Finally, there were no statistically significant differences between the results of the OCC techniques and those of logistic regression which could indicate the superiority of one approach over the other. Therefore, both approaches merit consideration when dealing with LDPs. Thomas (2009b) highlights the idea that a new methodology, using the same characteristics of the data as used by existing methods, producing a superior performance is questioned by many experts (see Hand, 2006a). Indeed, Overstreet *et al.* (1992) observe that based on the flat maximum

effect, the predictive performance of supposedly different classification techniques is almost indistinguishable as it is likely that most classification techniques will generate a model close to the best discrimination possible. Many other issues need to be considered when comparing the performance of a model, some of which have been outlined above and will be addressed in future work.

CHAPTER 6

Benchmarking Behavioural Scoring

For financial institutions, behavioural scoring is a valuable tool used to reduce loss and increase profit. This is achieved via the control of risk by assessing the ongoing creditworthiness and consumer behaviour of their existing customers. Behavioural scoring is used throughout the lifetime of the customers' relationship with the financial institution. With behavioural scoring models, not unlike application scorecard models, the dataset construction and modelling stages (e.g. Figure 3.3) are characterised by a range of key parameters settings pertaining to how the data should be evaluated. Because of the relative lack of authoritative evaluation studies published in the literature, the purpose of this chapter, by means of quantitative evaluation, is to address the following important questions:

- i. To what extent does the use of different durations of historical customer re-

payment data for model training affect the classification performance of a behavioural scoring model?

We evaluate the classification performance of various logistic regression models, each of which is trained using a particular duration of historical customer repayment data.

- ii. What impact do variations in the outcome window, from which a customer's class label is defined (good or bad), have on the classification performance of a behavioural scoring model?

From a practical perspective, the size of the outcome window is not an arbitrary decision made during the training of the classification model. Quite often specific business requirements (e.g. long term forecasts) determine the size of the outcome window. By varying the length of the outcome window, we quantify differences between performance results of various logistic regression models. Intuition would suggest a superior classification performance from models trained using a shorter outcome period compared to models constructed using a longer outcome period. The purpose of this assessment is to quantify such differences, so as to provide practitioners with a benchmark.

- iii. What are the differences between alternative approaches used to define a customer's default status?

We compare two approaches used to define the customers' default status. For practitioners, in some situations there are valid justifications for employing a specific label definition approach. For example, to alleviate the low-default portfolio problem, as examined in Chapter 5, practitioners may boost the

number of loan defaulters by adopting a particular label definition approach.

The purpose of this assessment is to quantitatively analyse the differences between both label definition approaches.

Guidance in the literature on the impact of choices made for key modelling parameters for behavioural scoring models are largely limited to anecdotal evidence that suggest good practices and processes for implementation. We address this shortcoming in the literature by conducting an empirical investigation into behavioural scoring. The findings reported in this chapter are based on real-world data from a credit bureau which details the performance of retail loans issued by the main Irish banks in 2003 and 2004. Overall, the purpose of the work described is to present a convenient source of comparison for practitioners who must make certain behavioural scoring modelling decisions. By identifying and quantifying the impact of different modelling decisions through the use of real-world data, clear insights into the dataset construction and modelling stages of behavioural scoring are gained.

Section 6.1 describes our experimental set-up, including the data used and the experimental methodology. In Section 6.2 we describe our experimental results. Finally, in Section 6.3 we reflect upon the implications of our findings for practitioners engaged in behavioural scoring.

6.1 Experiment Set-up

For the sake of clarity, material in Section 4.2.1 which is relevant to the experimental information is repeated below. In Figure 6.1 a sample of customers is selected so that their repayment behaviour either side of an arbitrarily chosen *observation point*

is available. The period before the observation point, which details a customer's historical repayment data, is often termed the *performance window*. Data on the customers' performance during this time is structured into features which are used by the behavioural scoring system to distinguish between customers' likely to repay their loan and those likely to default on their financial obligation.

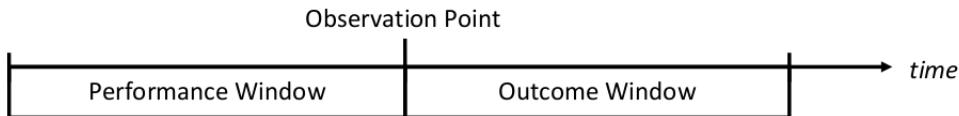


Figure 6.1: Behavioural scoring performance window and outcome window.

Previously, Section 3.2.1.4 described two approaches used to define the customers' default status: (i) the *worst status* label definition approach; and (ii) the *current status* label definition approach. The *current status* label definition approach classifies a customer as either good or bad based on their account status at the end of the outcome window. The *worst status* label definition approach classifies a customer as either good or bad based on their account status during the outcome window. For example, we define a bad as 2 or more missed loan repayments. If a customer's outcome window records the total number of missed loan repayments for the previous 6-months as [1, 2, 1, 0, 0, 1] then under the *current status* the customer is labelled as good. Using the *worst status*, the customer is labelled as bad.

The purpose of this evaluation is to investigate the efficacy of an assorted set of performance and outcome window sizes on the classification accuracy of logistic regression (LR) models in behavioural scoring scenarios. This is achieved by creating multiple behavioural scoring datasets with varying performance and out-

come window sizes. In this study three different performance window sizes are used (6-months, 12-months, and 18-months) along with 5 separate observation window sizes (3-months, 6-months, 12-months, 18-months, and 24-months). This study also assesses two popular approaches used to label behavioural scoring data, namely the *current status* label definition approach and the *worst status* label definition approach. To accomplish these aims we perform three overlapping experiments:

- i. A comparison of classifier performance under different window sizes by varying the performance window sizes over a range of fixed outcome window sizes.
- ii. A comparison of classifier performance under different window sizes by using a single fixed performance window size and varying outcome window sizes.
- iii. A comparison of classifier performance under different label definition approaches by using a single fixed performance window size and varying outcome window sizes, as per (ii).

This section describes the datasets, performance measure and methodology used.

6.1.1 Data

In these experiments we have used data provided by the Irish Credit Bureau (ICB). Credit bureau data is regarded by many in industry as a reliable source of data with the potential to increase productivity and profitability in credit decision making (Mays, 2004). This is due, in part, to: (i) the data privacy and consumer protection laws which help ensure that the data is highly regulated in terms of its use (Mays, 2004); and (ii) information sharing among lenders lowers credit risk. Typically, information in the ICB data is gathered from a number of different lenders who report

to the credit bureau every month. The ICB data details updates on customers' payment behaviour and any changes to public records, e.g. court judgements, collection items, and bankruptcy.

Previously, in Chapter 5, application scorecard models were constructed based on two separate snapshots of customer information - the first consisted of customer characteristics on applying for a loan and the second of customer default status sometime later (typically 12 months). In behavioural scoring the first snapshot is replaced by a broader aspect of customer behaviour collected over a duration of six to 24 months but the second snapshot remains the same.

In order to conduct the experiments described in this chapter, it was necessary to transform the ICB data into multiple experimental datasets. This data transformation process consisted of three steps: (i) dataset preparation; (ii) dataset generation; and (iii) dataset labelling. Dataset preparation, described in Section 6.1.1.1, involved collating and reviewing the ICB data. Dataset generation, described in Section 6.1.1.2, consisted of generating multiple datasets based on experimental parameters (e.g. performance window size). This step also included the specification and generation of additional features based on existing data. During the dataset labelling step, described in Section 6.1.1.3, a label definition approach was applied to the experimental datasets.

6.1.1.1 Dataset Preparation

The ICB data used in these experiments was anonymised to protect customer confidentiality and identity. It contains details of 2,500 customers who were approved for a mortgage loan between January 2003 and December 2004. The data provided

includes a subset of their application characteristics and full subsequent repayment behaviour up to December 2010. Based on statistics issued by the Irish government's Department of Environment, Community and Local Government (DofE, 2008), the number of Irish mortgages issued in 2003 stood at 97,726 (or €17,432m worth) and for 2004 the figure recorded was 104,134 (or €21,003m worth). As such, this data represents a random sample of just over 1% of all mortgages issued in Ireland between 2003 and 2004. Typically, for each customer, the ICB receive monthly data record updates; in this dataset each data record is updated every 3 months resulting in multiple data records per customer. Each data record details a customer's repayment behaviour for the previous 24-months from the time of the update.

Table 6.1 describes the features of each ICB data record. The features are grouped into customer loan application data (*Application data*) and customer repayment behaviour data (*Behavioural data*). The application data for each data record remains unchanged from the time of the original loan application. The behavioural data is updated every 3 months, as specified by *Account update date*. Depending on when data is received by the credit bureau from the banks, the account update date occurs either at the end of the third month or shortly afterwards at the beginning of the following month. *Loan Protection* is used to indicate if the customer has some form of financial protection against an unforeseen adverse personal event. The *Account association* feature indicates if the loan is associated with a single or joint account. The *Outstanding loan balance* is the overall amount owed. Also included is a non-standard ICB feature, *Loan installment amount*, which is the amount repaid since the last quarter. This is calculated as the difference between the current *Outstanding loan balance* and that of the previous quarter. Features that uniquely

identify the data record or customer were removed (see Table 6.1). Customers with missing or incomplete information were removed (18 in total).

Table 6.1: ICB data features. Features removed from the ICB data during the dataset preparation step are indicated by *.

Type	Variable Name	Description
Application data	Account opening date*	Date of loan drawdown
	Loan amount	Total repayment amount
	Term of loan	Length of loan in years
	Loan protection	Financial insurance
	Customer location	Current residence of the customer
	Account association	Single or joint account
	Payment frequency	Monthly or fortnightly repayment
Behavioural data	Date of birth*	Customer date of birth
	Account update date*	Date of loan update
	Loan installment amount	Amount repaid for update
	Outstanding loan balance	Overall amount owed
	Vintage	Age of loan in months
	Repayment indicators	Monthly loan status

Current credit bureau information is contained in the *Repayment indicators* feature. These features record not only the current status of the loan, but also the loan status for each of the previous 23 months. Standard credit bureau repayment indicators denote:

- Whether currently, or at any stage of the previous 23 months, the account has been closed (*closed*);
- The lender and customer have agreed to suspend all or part of the payment (*moratorium*);
- The account has been dormant¹ for a period of time (*dormant*);
- The account is pending litigation (*litigation*);

¹Borrowers sometimes retain a dormant account as security for another loan

- The account has been frozen by the lender (*frozen*);
- The account has no missed payments (*normal*);
- The account is in arrears for x number of repayments (*arrears*).

For example, Figure 6.2 illustrates the repayment indicator values for an account opened in February, during the first quarter of 2003. In total 11 separate data records (U_1 to U_{11}) detail the lifetime of the loan. The first account update date (U_1) was received by the credit bureau on the 30th of June 2003. This update details the account's repayment indicator feature values for each full month of loan repayments (i.e. March, April, May, and June of 2003). The next update, U_2 , was received by the bureau on the 6th of October 2003, and the account's repayment indicators feature values from March until September 2003. Subsequent updates, as far as U_{11} , were received close to the end of each quarter. At each update customer repayment behaviour for the previous 24-months is reported. If there are no missed loan repayments then 0 is used to indicate that the account is not in arrears. A repayment indicator between 1 and 9 represents the total number of months the customer is in arrears.

Based on Figure 6.2, according to U_4 the customer did not make a full loan repayment in January 2004, hence the 1 for January 2004. Similarly, a repayment indicator of 2 indicates that another repayment was missed in February 2004. The customer is now 2 months in arrears. The following month, March, the customer makes a full repayment for the month in addition to the outstanding amount for January. However, the missed repayment for February is still outstanding, hence a repayment indicator of 1 for March 2004. The C repayment indicator for U_{11} at

the end of December 2005 indicates that the customer account was closed and no outstanding amount was due from the customer. As a result no subsequent data records are received for this loan.

After collating and reviewing the ICB data, we obtained a dataset of close to 45,000 instances with each instance representing a quarterly snapshot of customer behaviour over the previous 24-months. The next step describes how the ICB data is transformed into multiple experimental datasets which are used to train and test the model.

6.1.1.2 Data Generation

In this step, based on the performance window size, multiple experimental datasets were generated using the ICB data. The performance window size dictates how many months worth of the repayment indicators are used to train and test the model. After the ICB data has been partitioned into multiple datasets based on the performance window size, additional feature values are generated using the repayment indicator feature values. Only a single instance of the quarterly snapshots of a customer's behaviour over the preceding 24-months is included in each dataset. The most recent instance (i.e. the one closest to the performance window end date) is retained. The remainder of this section describes the process used to generate these additional feature values.

In order to extract the maximum amount of information from the data, an additional set of features are derived from the repayment indicators to form the *combination features*. The combination features describe the state of the account over the performance window. The ICB data features (see Table 6.1) in addition to the com-

Figure 6.2: Example of a customer account detailed by 11 data records.

bination features are included in the experimental datasets. Depending on the size of the performance window, a maximum of 61 combination features are defined. For example, Table 6.2 lists the combination features derived from the *arrears repayment indicator*. The combination features for the remaining repayment indicators are derived in the same manner. Table 6.3 to Table 6.8 detail the remaining repayment indicators.

Table 6.2: Combination features generated based on the *arrears repayment indicator*. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.

<i>Feature Name</i>	<i>Description</i>
Arrears ever	Have arrears occurred at any stage during the performance window
Arrears 1-to-3	Have arrears occurred in the last 3 months
Arrears 4-to-6	Have arrears occurred in months 4-to-6
Arrears 7-to-12*	Have arrears occurred in months 7-to-12
Arrears 13-to-PWSize*	Have arrears occurred in month 13 or later
Arrears worst	Highest number of missed repayments
Missed single	Only one missed repayment on a single occasion
Missed twice	Only one missed repayment on two separate occasions
Missed thrice	Only one missed repayment on three separate occasions
Missed multi	Only one missed repayment on more than three separate occasions
Missed double	At most 2 repayments in arrears
Missed treble	At most 3 repayments in arrears
Missed cont > 3	At most 3 or more repayments in arrears
Arrears closed	Was the account in arrears before it was closed
One current	Has the account gone directly from 1 missed repayment to normal
Two current	Has the account gone directly from 2 missed repayments to normal
Three current	Has the account gone directly from 3 missed repayments to normal

The arrears repayment indicator is used to create features to indicate the following: that the account has been in arrears sometime during the performance window; that arrears have occurred over certain time frames (e.g. in the last 3-months); the highest number of missed payments; that only one missed payment has occurred; that a missed payment has occurred on two separate occasions; and that the account has moved from arrears to current (i.e. no outstanding arrears). The size of the performance window (*PWSize*) may preclude certain features from being used (e.g.

if PWSize = 12-months, then the Arrears 13-to-PWSize feature will not be used).

Table 6.3: Combination features generated based on the *normal* repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.

<i>Feature Name</i>	<i>Description</i>
Normal at present	Is the account currently Normal
Normal always	Has the account been Normal for the entire performance window
Normal 1-to-3	Has the account been Normal for the previous 3 months
Normal 4-to-6	Has the account been Normal for months 4-to-6
Normal 7-to-12*	Has the account been Normal for months 7-to-12
Normal 13-to-PWSize*	Has the account been Normal from month 13 and onwards
Normal < 12*	Has the account been Normal for all of the previous 12 months

Table 6.4: Combination features generated based on the *moratorium* (Morat.) repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.

<i>Feature Name</i>	<i>Description</i>
Morat. ever	Has a moratorium occurred during the performance window
Morat. 1-to-3	Has the account been in moratorium in the last 3 months
Morat. 4-to-6	Has the account been in moratorium in months 4-to-6
Morat. 7-to-12*	Has the account been in moratorium in months 7-to-12
Morat. 13-to-PWSize*	Has the account been in moratorium in month 13 or later
Most Morat.	What is longest the account has been in moratorium for
Total Morat.	The total number of months the account has been in moratorium
Arrears-to-Morat.	Has the account been in arrears immediately prior to moratorium

It was necessary to exclude a number of instances from the experimental datasets used to train and test the model. Any instances less than 3 months old (measured using the account opening date and the performance window end date) were removed (Mays, 2004, pp.151). Instances marked as closed from the beginning of the performance window were also removed.

This section has described how the ICB data was transformed, based on the performance window size, into multiple experimental datasets. The process of generating an additional set of features, known as combination features, was also detailed. The next step describes how a class label is applied to the instances of the

Table 6.5: Combination features generated based on the *dormant* repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.

<i>Feature Name</i>	<i>Description</i>
Dormant ever	Has a dormancy occurred during the performance window
Dormant 1-to-3	Has the account been dormant in the last 3 months
Dormant 4-to-6	Has the account been dormant in months 4-to-6
Dormant 7-to-12*	Has the account been dormant in months 7-to-12
Dormant 13-to-PWSize*	Has the account been dormant in month 13 or later
Most Dormant	What is longest the account has been dormant for
Total Dormant	The total number of months the account has been dormant
Arrears-to-Dormant	Has the account been in arrears immediately prior to dormancy

Table 6.6: Combination features generated based on the *litigation* (Litig.) repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.

<i>Feature Name</i>	<i>Description</i>
Litig. ever	Has litigation occurred at any stage during the performance window
Litig. 1-to-3	Has the account been subject to litigation in the last 3 months
Litig. 4-to-6	Has the account been subject to litigation in months 4-to-6
Litig. 7-to-12*	Has the account been subject to litigation in months 7-to-12
Litig. 13-to-PWSize*	Has the account been subject to litigation in month 13 or later
Most Litig.	What is longest the account has been subject to litigation for
Total Litig.	The total number of months the account has been subject to litigation
Arrears-to-Litig.	Has the account been in arrears immediately prior to litigation

Table 6.7: Combination features generated based on the *frozen* repayment indicator. PWSize = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size.

<i>Feature Name</i>	<i>Description</i>
Frozen ever	Has the account been frozen at any stage during the performance window
Frozen 1-to-3	Has the account been frozen in the last 3 months
Frozen 4-to-6	Has the account been frozen in months 4-to-6
Frozen 7-to-12*	Has the account been frozen in months 7-to-12
Frozen 13-to-PWSize*	Has the account been frozen in month 13 or later
Most Frozen	What is longest the account has been frozen for
Total Frozen	The total number of months the account been frozen

Table 6.8: Combination features generated based on the *closed* repayment indicator. PWSIZE = performance window size in months. * indicates that the inclusion of the feature depends on the performance window size. Note: Once Closed, the loan status remains unchanged.

<i>Feature Name</i>	<i>Description</i>
Closed 1-to-3	Was the account closed in the previous 3 months.
Closed 4-to-6	Was the account closed in months 4-to-6.
Closed 7-to-12*	Was the account closed in months 7-to-12.
Closed 13-to-PWSIZE*	Was the account closed in month 13 or later
Arrears-to-Closed	Was the account in arrears immediately before it was closed.
Morat-to-Closed	Was the account on moratorium immediately before it was closed.

experimental datasets.

6.1.1.3 Dataset Labelling

The instances in the generated experimental datasets need a class label in order to build a model. The class label indicates whether or not the customer defaults on their loan obligation. Normally, as per Basel II, a default occurs when the customer is past due on a loan for 90-days or more, i.e. three or more missed monthly payments. With the ICB data, when a customer is defined as bad using this approach too few defaulters are generated with which to train the LR model, i.e. an extreme version of the low-default portfolio problem examined in Chapter 5. In these experiments a customer is defined as bad if they are more than 60-days in arrears on their loan, i.e. two or more missed monthly payments. Figure 6.3 displays the monthly default rate (i.e. customer accounts with 2 or more missed payments) for each month from January 2003 to December 2010. The monthly default rate for customer accounts opened in 2003 and 2004 are detailed. A possible explanation for the spike in the default rate experienced by loans issued in 2004 can be attributed to higher property prices and larger loan amounts. Given the low default rate as evidenced by Figure

6.3, the ICB dataset can be considered as representative of the low-default portfolio problem discussed in Chapter 5.

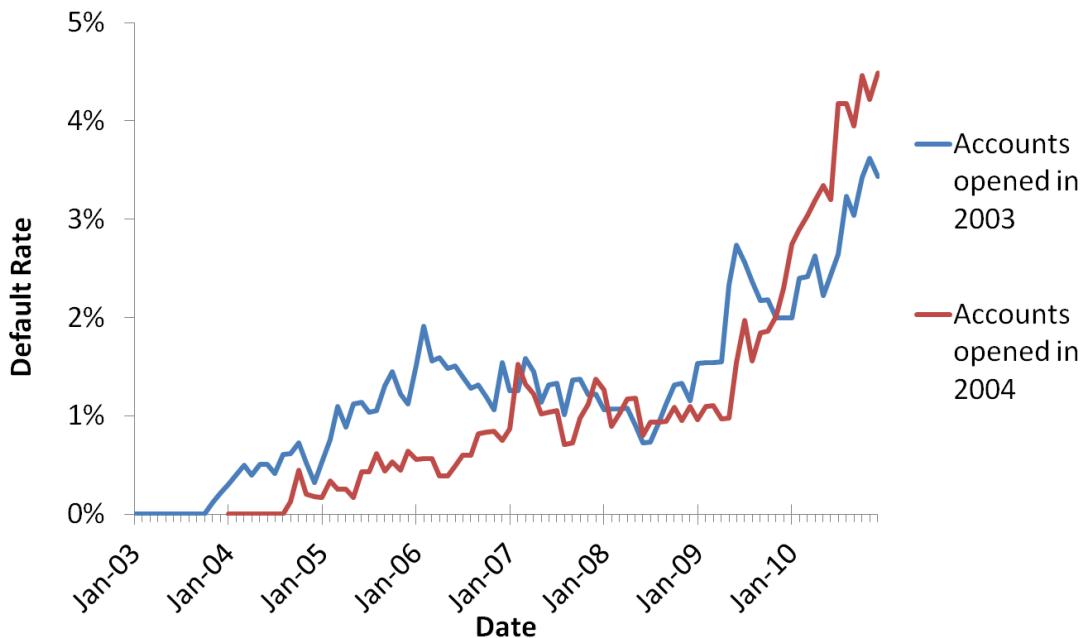


Figure 6.3: Monthly default rate, up until December 2010, for customer accounts opened in 2003 and 2004.

In order to maximise the separation between the goods and the bads, indeterminates were omitted from the data used to train the model. An indeterminate was defined as someone whose behaviour is somewhere between good and bad, i.e. 1 month in arrears (Thomas *et al.*, 2001b). Indeterminate accounts were retained in the test data.

The next section describes the performance measure used in these experiments.

6.1.2 Performance Measures

In Chapter 5 classifier performance was assessed using the harmonic mean and the H measure. In the context of the low-default portfolio problem, an attractive quality of the harmonic mean is that it is less sensitive to large numeric values and more

sensitive to small values (Butler *et al.*, 2009). With low-default portfolios, such small values can often occur when measuring the classification performance of the minority class. Thus, we justify using the harmonic mean based on the relatively poor discriminability of certain datasets, e.g. PAKDD, Spain, and Thomas. Generally, it is accepted that behavioural scoring is more accurate than application scoring (Hand, 2001). Therefore, we dispense with the harmonic mean and select, in the interests of relevancy, the more widely used and straight forward average class accuracy (see Equation 2.24). A validation dataset (see Section 6.1.3) was used to identify an optimised threshold at which to make actual classifications. This is in keeping with the approach used in Chapter 5, which highlighted how the low-default portfolio problem can be addressed by optimising the threshold on classification output.

Differences between the performance of the LR classifier on various test datasets (see Section 6.1.3) were analysed with a Kruskal-Wallis one-way analysis of variance by ranks test (see Wolfe & Hollander, 1999) and *post hoc* pairwise comparisons performed with a Dwass-Steel-Critchlow-Fligner procedure (see Wolfe & Hollander, 1999). The Kruskal-Wallis one-way analysis of variance by ranks test (or simply, Kruskal-Wallis test) is a non-parametric method for testing whether k groups have been drawn from the same population. If the result of the Kruskal-Wallis test is significant, it indicates that there is a significant difference between at least two of the groups in the set of k groups (Sheskin, 1997). As the test is non-parametric, it makes no assumption about the shape of the population distribution from which the groups are drawn. The Kruskal-Wallis is used because an unequal number of observations were recorded for the LR classifier on various test datasets. In Chapter 5, a Friedman test was used as an equal number of observations (i.e. classifier performance) were

obtained for each group (i.e. class imbalance ratio). When significant differences were indicated by the Kruskal-Wallis test, all possible pairwise comparisons between the groups were made using the Dwass-Steel-Chritchlow-Fligner *post hoc* test to identify specific group differences.

A two-tailed Mann-Whitney U test (see Wolfe & Hollander, 1999) was used to test for statistically significant differences between the average class accuracies of the *worst status* and *current status* label definition approaches. For all statistical tests, significance was established at $p < 0.05$ and statistical analysis was performed using the statistical package StatsDirect (Buchan, 2011). The Mann-Whitney U test is appropriate for comparison of two groups only, unlike the Kruskal-Wallis test which is the most appropriate test when all sample groups are compared with each other.

The next section describes how the multiple experimental datasets are used to train and test a classification model.

6.1.3 Methodology

Training data from the experimental datasets was comprised solely of customer accounts opened in 2003. The training data was then divided into two subsets: (i) the classifier training set (67%); and (ii) the validation set (33%). The classifier training set and the validation set were used to train and tune the classifier. Test sets from the experimental datasets were used to evaluate classifier performance and were comprised solely of customer accounts opened in 2004. This ensured that the customer accounts in the training and test sets were mutually exclusive. The performance window start date was set as the beginning of January 2005 for all of the

initial test sets. For subsequent test sets the performance and outcome window start dates were incremented by 3-months, e.g. for the second test set the performance window start date moved to the beginning of April 2005. Classifier performance was measured and the process was repeated until the outcome window end date reached the end of 2010.

Figure 6.4 displays the experimental set-up for a 12-month performance window and 3-month outcome window. For the training data, the performance window start date was set as the beginning of January 2004. The performance window end date, determined by the performance window size, was set as the end of December 2004. The outcome window began immediately after the performance window end date and its end date, determined by the outcome window size, was set as the end of March 2004.

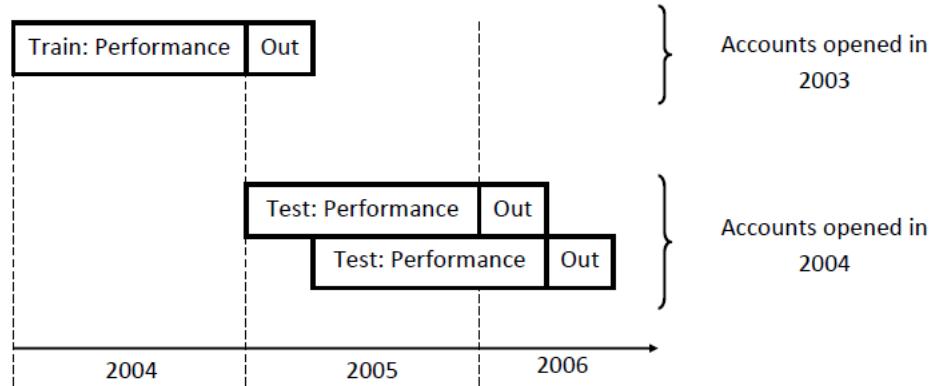


Figure 6.4: Experiment set-up for a 12-month performance window and a 3-month outcome window (Out).

The first stage of our evaluation, (i), considers the classification performance of a LR model constructed using a selection of performance window sizes. The comparison is performed using the *worst status* label definition approach. Based on the results of (i) the most accurate performance window size is then used in the

next stage of the analysis, (ii), whereby the outcome window sizes are compared using both the *worst status* approach and the *current status* approach. The final stage of the analysis, (iii), performs a direct comparison of the *current status* and *worst status* label definition approaches. The comparison is performed using the datasets generated for (ii). In the study the performance window sizes used are 6-months ($P6$), 12-months ($P12$), and 18-months ($P18$). The five outcome window sizes used are 3-months ($O3$), 6-months ($O6$), 12-months ($O12$), 18-months ($O18$), and 24-months ($O24$).

6.1.4 Model Training

As previously highlighted in Section 6.1.1, the real-world dataset used in this study resembles the low-default portfolio problem. By extending the findings presented in Chapter 5 to behavioural scoring, at such low levels of default, one should consider using either a LR model or a semi-supervised one-class classification technique with which to construct a behavioural scoring model. In this work a LR model was trained and tested on our experimental datasets for a number of reasons. Firstly, LR was selected as it is widely used to build credit scoring models (Hand & Zhou, 2009). As highlighted at the beginning of this chapter, the purpose of this study is to provide practitioners with a source of comparison for behavioural scoring and so it is most appropriate to persist with the most commonly used modelling approach. Secondly, despite the existence of more sophisticated classification models such as support vector machines and neural networks, the popularity of LR has endured. This may be due to the interpretability and fast estimation of its parameters, or, as demonstrated in Chapter 5, the favourable classification performance achieved by

LR when compared to other supervised classification approaches.

The LR model was implemented using the Weka (version 3.7.1) machine learning framework (Witten & Frank, 2005). Tuning involved using the validation set with which to optimise the LR ridge estimator parameter in order to offset unstable coefficient estimates that arise from highly correlated data or when the number of features is relatively large compared to the number of observations. Miller (1984) describes ridge regression as a *shrinkage technique* that can be used as an alternative to other established feature selection approaches, some of which are described in Section 3.2.2.1. The ridge estimator parameter is used to overcome overfitting and collinearity by specifying a restriction on the coefficients, β , used by logistic regression. This restriction reduces the bias of the regression coefficients. As a drawback, using such an approach can lead to an over-parametrised model that can be difficult to interpret and maintain. Each experiment was conducted 10 times using different randomly selected training and validation set splits, and the results reported are averages of these 10 runs.

Section 6.2 will describe the results of this experimental process.

6.2 Results and Discussion

The results and findings reported in this quantitative evaluation are intended to serve as a source of comparison for practitioners conducting similar assessments, which may then assist practitioners in building accurate behavioural scoring models. As such, the objectives of this evaluation were threefold:

- i. First, by adjusting the performance window size, we compare the classification

results of LR models trained using different durations of historical customer repayment data. This comparison is repeated for each of the five outcome window sizes ($O3$, $O6$, $O12$, $O18$, and $O24$). The findings reported in this evaluation provide guidelines on what duration of historical customer repayment data should be used when attempting to accurately predict loan defaults within a specified time frame.

- ii. The LR model yielding the best overall average class accuracy is passed to the next stage of the evaluation. In this stage, we quantify differences between LR model performance results based on different outcome window sizes. Intuitively, a superior classification performance is expected from LR models constructed using class labels obtained over a short outcome window compared to models constructed using class labels obtained from a longer outcome window. The purpose of this assessment is to quantify such differences. This is an important topic for practitioners, as business requirements (e.g. short term forecasts) can often dictate the size of the outcome window.
- iii. Finally, we compare two approaches used to define the customers' default status: (i) the *worst status* label definition approach; and (ii) the *current status* label definition approach. Similar to (ii), we quantify differences between LR model performance results for each outcome window size. The LR models are trained using a 12-month performance window and a particular label definition approach. To the best of our knowledge no such evaluation has been published in the behavioural scoring literature. This evaluation offers practitioners a source to compare and to appraise differences between both label definition

approaches. Ultimately though, as per the outcome window, business and regulatory requirements and can often determine which label definition approach to use.

6.2.1 Performance Window Selection

For each of the five separate outcome window sizes, Figure 6.5 provides a graphical representation of the LR models' classification results when a particular performance window is used. The reported results are based on the *worst status* label definition approach. Beyond a 6-month outcome window the models' average class accuracies decrease as the size of the the outcome window increases. By way of comparison, Figure 6.6 illustrates the *current status* label definition approach. The patterns of both figures are broadly similar, though mismatches exist for both the 6-month and 12-month outcome windows. As Basel II adopts the *worst status* label definition approach, and for clarity, the results described in the remainder of this subsection focus exclusively on this approach.

From Figure 6.5, a number of general points are apparent. Firstly, the LR model using the 12-month performance window outperformed both of the other models (P6 and P18) when used with shorter outcome windows (i.e., O3, O6, and O12). This is interesting as it highlights how the 6-month performance window is too short to allow for a sufficient accumulation of transitions in customer repayment behaviour with which to build a stable classification model. Conversely, the 18-month performance window may be too long and events occurring earlier in the performance period should not be given the same weight or importance as more recent transitions. Finally, the average class accuracy of the LR model using the

6-month performance window frequently did not achieve parity with models using the longer performance windows. This relates to the previous point regarding the stability of the classification model.

For each different outcome window size used in this experiment differences in classifier performance were compared using the Kruskal-Wallis test for significance. The results of the tests established at least one significant difference between the results of the performance windows in each of the outcome window categories, except for O24. For O24 no significant difference between the average class accuracies of the three performance window sizes was detected, which is unsurprising considering Figure 6.5e. Table 6.9 displays the results of the Dwass-Steel-Chritchlow-Fligner *post hoc* test for pair-wise comparisons.

Table 6.9: Average class accuracy *post hoc* analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for the *worst status* (worst) label definition approach are provided. Note, no statistical significance was detected between the average class accuracies using O24. Statistical significance is indicated by *.

<i>Outcome</i>	<i>Performance</i>	<i>p-value</i>
O3	P18 vs. P12	0.0006*
	P18 vs. P6	0.0001*
	P12 vs. P6	< 0.0001*
O6	P18 vs. P12	0.0003*
	P18 vs. P6	0.5318
	P12 vs. P6	< 0.0001*
O12	P18 vs. P12	0.0041*
	P18 vs. P6	0.9919
	P12 vs. P6	0.0003*
O18	P18 vs. P12	0.4409
	P18 vs. P6	0.6139
	P12 vs. P6	0.0085*

On each of the shorter outcome windows (O3, O6, and O12) statistically significant differences between P12 and both of the other performance windows (P6 and

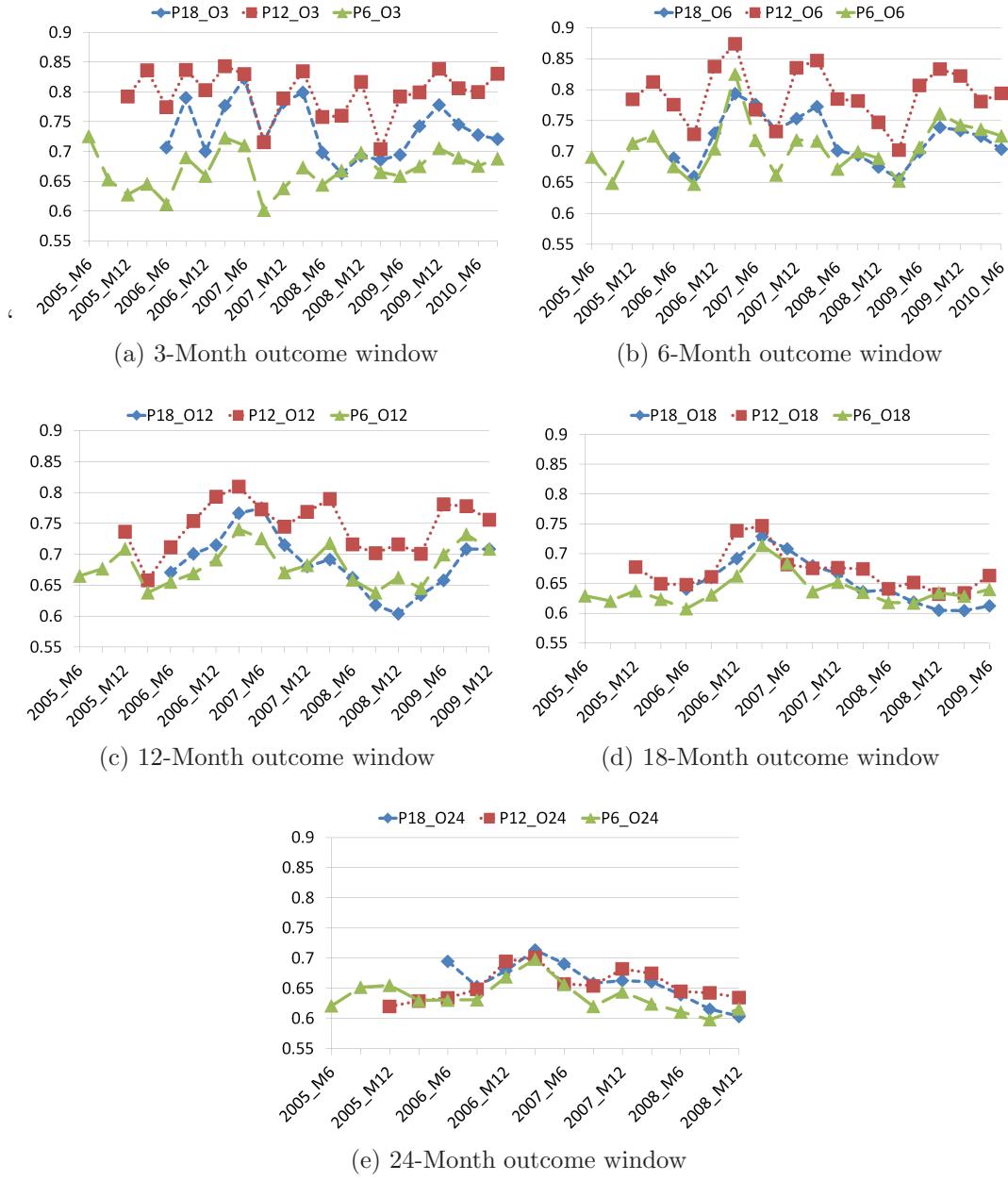


Figure 6.5: Average class accuracies (y -axis) of the behavioural scoring classification model when a particular combination of performance window and outcome window size definitions are used. The *worst status* label definition in all cases.

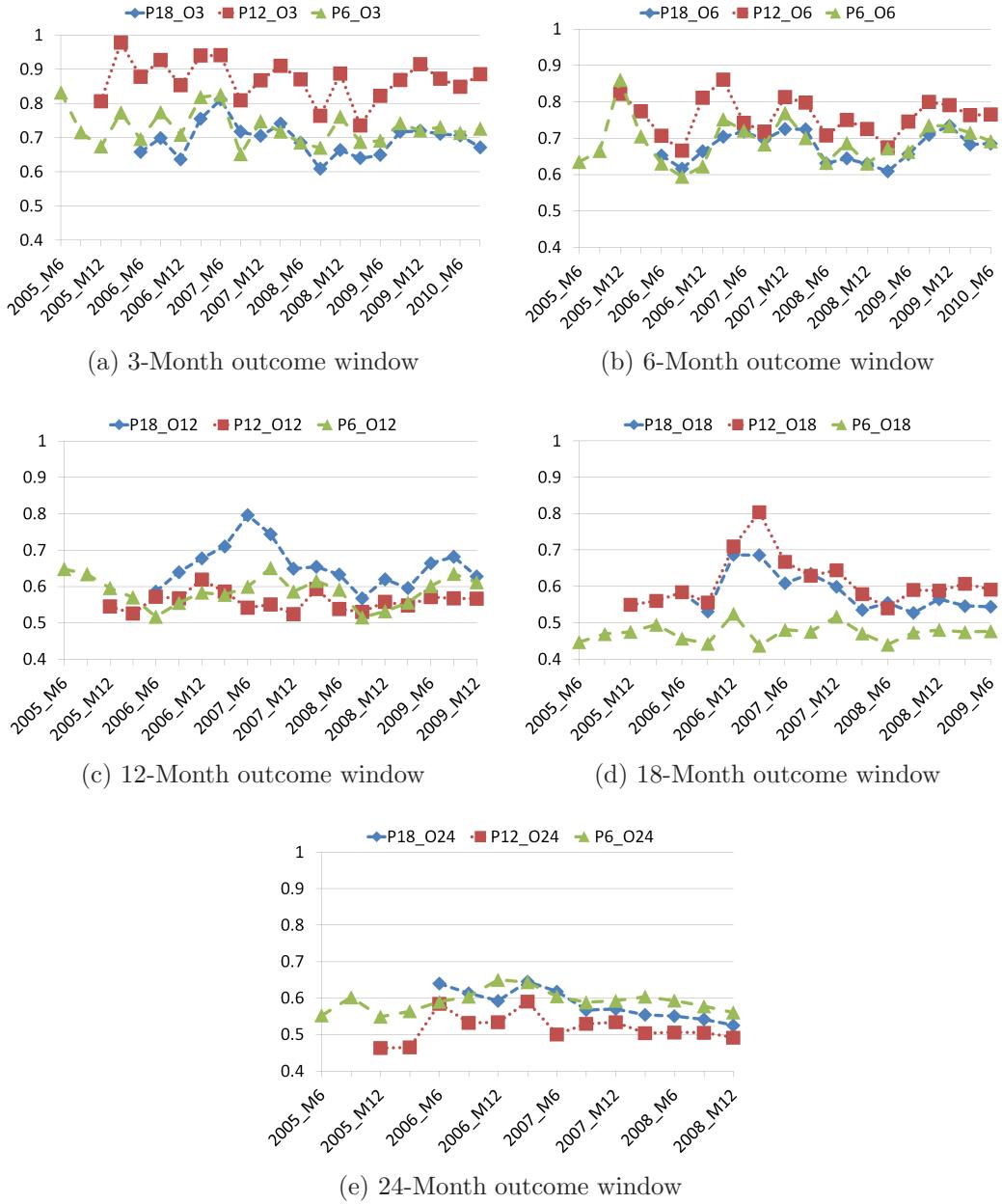


Figure 6.6: Average class accuracies (y -axis) of the behavioural scoring classification model when a particular combination of performance window and outcome window size definitions are used. The *current status* label definition in all cases.

P18) were detected. This suggests that a LR model trained with data based on a 12-month performance window is the most suitable for similar sized, or shorter, outcome windows.

When the outcome window is extended to 18-months, statistically significant differences in the average class accuracies of P6 and P12 were detected. When considered together with Figure 6.5d, this suggests that when using an 18-month outcome window, a LR model trained with data based on a 6-month performance window is the least suitable. No significant difference existed between the average class accuracies of P12 and P18. This once again underlines how the lack of accumulated customer repayment behaviour transitions affect classifier performance.

To summarise, the performance of three separate classification models, each one trained with data based on a particular performance window size, were compared using each of the five possible outcome window sizes. We consider a LR model trained with data based on a 12-month performance window as best suited to the classification task - particularly when outcome window sizes of 3-months, 6-months, and 12-months were specified. The literature [e.g. Thomas *et al.* (2001b)] generally recommends a performance window of 6 to 12 months. Our work supports this recommendation and further demonstrates, through quantitative evaluation, how the classification performance of a LR model trained with data based on a 12-month performance window is significantly better than its alternatives (P6 and P18) for three of the five outcome window sizes (O3, O6, and O12). Whilst our findings are context specific, we believe they offer a robust foundation for the understanding and development of future behavioural scoring applications. For the longer outcome windows sizes (O18 and O24), no statistical significance between the average class

accuracies of the classification models were detected to indicate the most suitable performance window. This is an interesting result as it suggests, regardless of performance window size, a plateau in classification performance when attempting to identify future loan defaults beyond a certain point in time. The LR model trained with data based on a 12-month performance window is carried forward into the next section where we quantify the effect varying outcome window sizes have on classification performance.

6.2.2 Outcome Window Selection

In this section we quantify the effects of adjusting the outcome window size on the average class accuracy of a LR classifier. As discussed in Section 4.2.1, the literature offers no clear consensus on the most appropriate outcome window size. We compare differences between LR models trained with data based on a 12-month performance window and a range of different outcome window sizes. As we are examining the impact on classifier performance caused by adjusting the size of the outcome window, it is important that these comparisons are performed using both the *worst status* and *current status* label definition approaches.

Figure 6.7 and Figure 6.8 compare classifier performance when different outcome window sizes are used together with a fixed performance window size of 12-months, using the *worst status* and *current status* approach, respectively. A cursory glance at both figures reveals, as expected, that the classification performance of the LR models using the shorter outcome windows (i.e. O3 and O6) is superior to those using longer outcome windows (i.e. O18 and O24).

In examining Figure 6.7, it is clear that a divide between the performance of the

classifiers built based on the shorter outcome windows (O3, O6, and O12) and those built based on the longer outcome windows (O18, O24) exists. Also, the average class accuracies of the LR models built based on 3-month and 6-month outcome windows appear to be closely correlated.

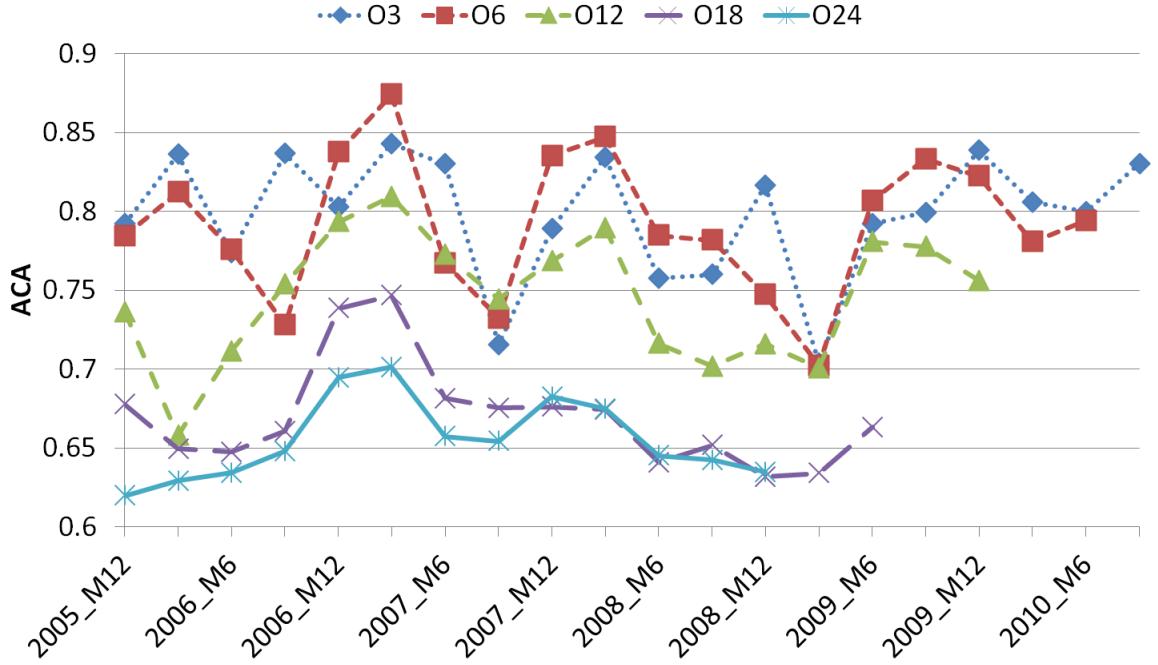


Figure 6.7: Average class accuracy (ACA) comparison of LR models using 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. *Worst status* label definition. The performance window is fixed at 12-months.

Figure 6.8 reveals a clear ordering of classifier performance relative to the size of the outcome window. A LR classifier trained with data based on a 3-month outcome window consistently achieves the highest average class accuracy followed by, in order, LR classifiers using O6, O18, O12, and O24. The relatively ineffectual performance of the LR classifier using the longest outcome window, 24-months, seems to suggest that the data is unsuited to predicting loan defaults occurring over a large outcome window size.

Both Figure 6.7 and Figure 6.8 confirm the intuitive expectation that classi-

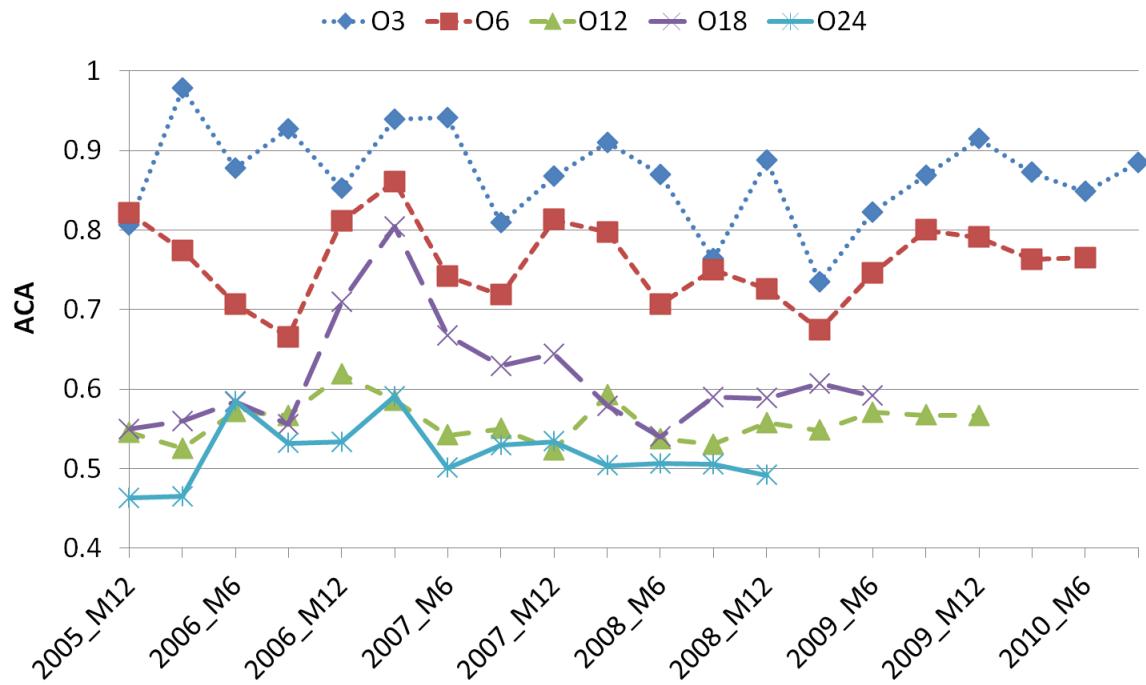


Figure 6.8: Average class accuracy (ACA) comparison of 3-month, 6-month, 12-month, 18-month, and 24-month outcome windows. *Current status* label definition approach. The performance window is fixed at 12-months.

fication accuracy degrades over longer outcome windows. Quite often though, as discussed in Section 4.2.1, it is business requirements (e.g. lender policies) and data constraints (e.g. sufficient number of defaulters) that determine the size of the outcome window. Therefore, as a source of comparison, it is important to measure the precise extent to which outcome window size influences classifier performance.

To determine differences between the average class accuracies of the LR models using the *worst status* definition, a multiple-comparisons test on all rank sums was conducted using the Kruskal-Wallis test. The results of the test established at least one significant difference between the results. Similarly, a Kruskal-Wallis test for significance on classifier performance when the *current status* definition is used established at least one significant difference between the average class accuracies. Table 6.10 displays the results of the Dwass-Steel-Chritchlow-Fligner *post hoc* test

for pair-wise comparisons for both the *current status* and *worst status* approaches.

Table 6.10: Average class accuracy of LR models trained with data based on a 12-month performance window, *post hoc* analysis of Kruskal-Wallis test using Dwass-Steel-Chritchlow-Fligner. Results for both label definition approaches: *worst status* (worst) and *current status* (current) are provided. Statistical significance is indicated by *.

Outcome	p-value	
	worst	current
O3 vs. O6	0.9766	0.0001*
O3 vs. O12	0.0052*	< 0.0001*
O3 vs. O18	< 0.0001*	< 0.0001*
O3 vs. O24	< 0.0001*	< 0.0001*
O6 vs. O12	0.0369*	< 0.0001*
O6 vs. O18	< 0.0001*	0.0002*
O6 vs. O24	< 0.0001*	< 0.0001*
O12 vs. O18	0.0006*	0.0394*
O12 vs. O24	< 0.0001*	0.0263*
O18 vs. O24	0.8156	0.0011*

When the *worst status* approach is used, no statistical significance was detected between the performance of LR models using O3 and O6. However, statistically significant differences were found between the average class accuracies of LR models using both O3 and O6 when compared with the classification performance of LR models using O12, O18, and O24. Similarly, statistically significant differences were detected between the performance of a LR model trained with data based on a 12-month outcome window and that of both O18 and O24. These results indicate 3 groups, LR models using O3 and O6, which are better than O12, which, in turn, is better than O18 and O24. This suggests that, based on the relatively superior average class accuracies of LR models using O3 and O6, shorter outcome windows are best suited to a LR model which uses a 12-month performance window for the current classification task. No statistical significance was detected between the

average class accuracies of LR models using O18 and O24. It is worth noting that the performance of a classifier degrades substantially when the outcome window is extended beyond a 12-month window. For example, the difference in classifier performance between a 6-month and 12-month outcome window is relatively less than the difference in classifier performance between a 12-month and 18-month outcome window.

When analysing the *current status* approach, statistically significant differences between the average class accuracies of each outcome window were detected. Based on classifier performance trained with data based on a 12-month performance window, this then enables us to infer the order of the most suitable outcome window size as follows: 3-months, 6-months, 18-months, 12-months, and 24-months. Note that the performance of a classifier degrades substantially once the outcome window is extended beyond 6-months. In contrast, for the *worst status* approach this degradation occurs when the outcome window is extended beyond 12-months.

To summarise, the results of the statistical comparison between the outcome windows clearly indicate that a LR classifier using either a 3-month or 6-month outcome window, in conjunction with a 12-month performance window, results in a significantly higher average class accuracy than that of a classifier using a longer outcome window size. This would suggest that the data does not contain information which allows a LR classifier to reliably distinguish between loan repayers and defaulters beyond a 6-month outcome window. For the *current status* approach, the separation between the performance of the LR classifiers is more pronounced compared to the *worst status* approach. In the next subsection, the reasons for such differences are

further explored.

6.2.3 Current Status versus Worst Status

This section performs a direct comparison of the classification performance of behavioural scoring models trained using either a *current status* or *worst status* label definition approach. Sometimes, by default, banks select a particular label definition approach due to a shortage of history on arrears or because of regulatory requirements. The purpose of this comparison is to provide practitioners with a published benchmark. In addition, as highlighted in Section 6.2.2, differences in both approaches are investigated further. The results of the comparison, performed with a LR model trained with data based on a 12-month performance window in combination with one of five outcome windows considered in Section 6.2.2, are illustrated in Figure 6.9. After the 6-month outcome window the performance of the LR model using the *current status* approach degrades rapidly, to the point where the average class accuracy is less than 0.5 when a 24-month outcome window is used. It is clear from Figure 6.9 that a higher average class accuracy can be achieved for a classification task based on the *worst status* approach and over longer outcome window sizes. However, a LR model using the *current status* approach scored higher using the 3-month outcome window. A two-tailed Mann-Whitney U test was conducted for each outcome window to find statistically significant differences between the *current status* and *worst status* label definition approaches. At this point some caution in interpreting quantitative differences is warranted. Is it reasonable for such significant differences in customer repayment behaviour to occur over such a short outcome window of 3-months? The differences in classifier performance may

be an artefact of the data. Despite this caveat, a broad pattern to the results of the comparison begins to emerge.

For all outcome windows, except the 6-month outcome window, the two-tailed Mann-Whitney U test found statistically significant differences between the average class accuracies of the LR models using the *current status* approach and *worst status* approach. This suggests that a LR model using the *worst status* approach with an outcome window of 12-months or longer has greater accuracy at identifying defaulters than when the *current status* approach is used. A LR model trained with data based on a 3-month outcome window, the *current status* approach presents a classification task that can be performed more accurately compared to the *worst status* approach.

The *worst status* approach uses the entire outcome window and even though a customer may recover from arrears at some point during the outcome window, they are still classed as bad. This may be problematic as it implies that the relative rankings of default risk hold for the entirety of the outcome window (Thomas *et al.*, 2001b). Conversely, with the *current status* approach, a customer who recovers from arrears during the outcome window is classed as good. Although a LR model using the *worst status* approach achieves greater accuracy using longer outcome windows, it may not recognise recent improvements to customers' financial circumstances. Overall, it is reasonable to assume that the data used by the performance window will contain information about whether or not a customer is likely to default on their loan. However, it is unlikely that a customer's behavioural patterns will indicate whether or not the customer will emerge from arrears.

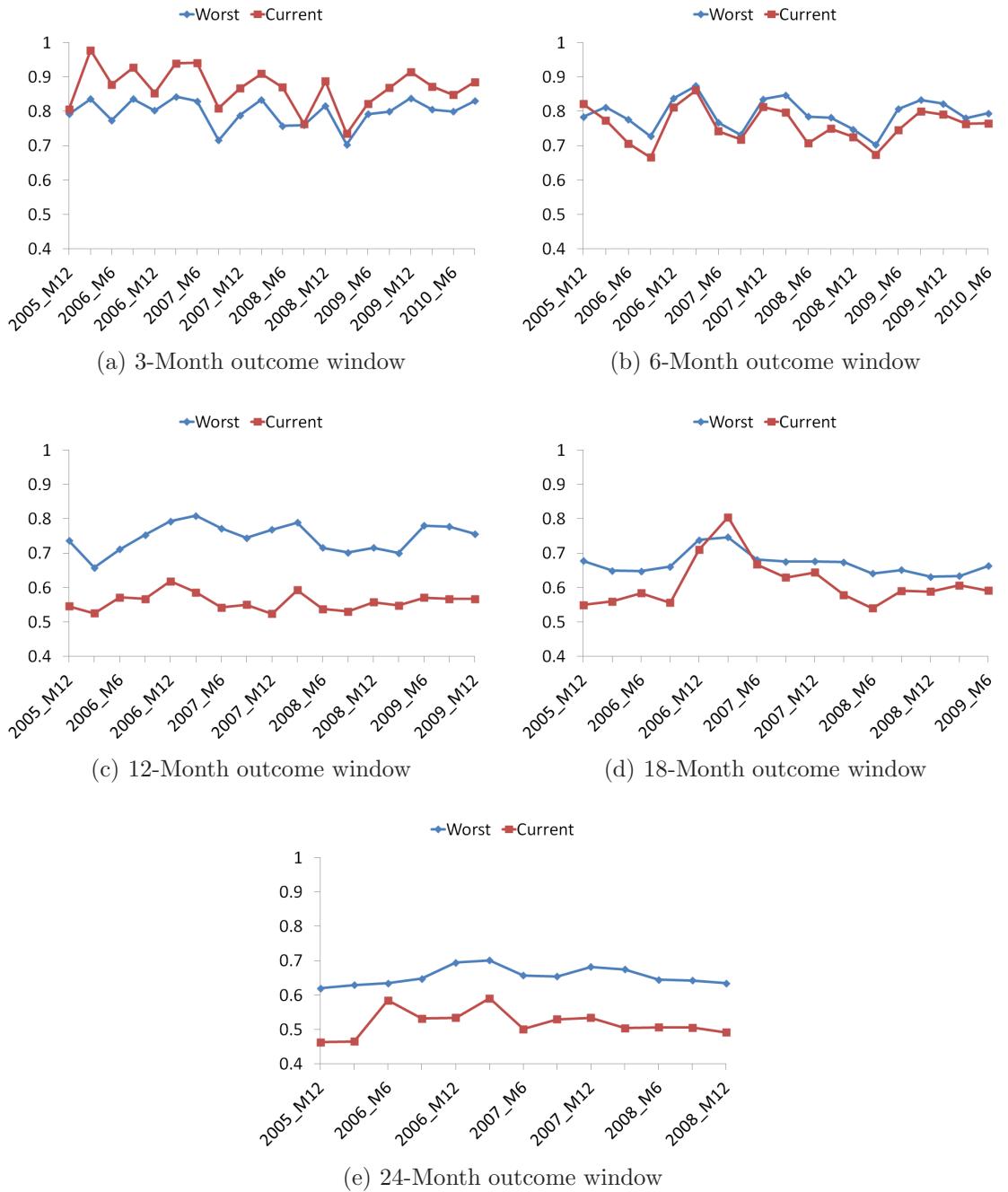


Figure 6.9: LR model average class accuracies (y -axis) trained with data based on a 12-month performance window and each outcome window. *Current status* versus *worst status*.

6.3 Conclusion

Behavioural scoring allows lenders to assess the likelihood of customers defaulting on their obligation during some specific outcome window. This assessment is based on customers' repayment behaviour over a fixed performance window. A customer's class label is defined either at the end of the outcome period (*current status*) or during the outcome period (*worst status*).

Using 7-years worth of data from the Irish market this chapter presented an extensive evaluation of behavioural scoring models built using varying performance and outcome time horizons. In addition, the chapter also detailed an empirical comparison of both label definition approaches.

Of the three separate performance windows used, a LR model using the 12-month performance window reported the highest average class accuracy when used with the shorter outcome window sizes. Over longer outcome window sizes the results exhibited greater ambiguity making it less discernible to identify an optimum performance window size. Frequently though, the 6-month performance window was unable to match the performance of the 12-month and 18-month performance windows. This relatively poor performance may be attributed to the fact that less data is used to train the model.

The impact of outcome window size was examined using a 12-month performance window. Rather predictably, of the 5 five different outcome window sizes, a LR model trained with data based on a 3-month outcome window resulted in the best accuracy. Quite often the gap in performance between LR models using either a 3-month outcome window or 6-month outcome window was statistically

insignificant. As the length of the outcome windows increased, the differences in the average class accuracy of LR models using the various performance windows became less distinguishable. Based on these results, it is reasonable to suggest that credit bureau data, alone, does not contain enough information with which a LR classifier can reliably distinguish between loan repayers and defaulters beyond a 6-month outcome window. Quite often the outcome window size is determined by business considerations or requirements. Such requirements need to consider short term forecasting.

The evaluation of the behavioural scoring models was conducted using two separate label definition approaches. LR models using the *worst status* approach performed better than the *current status* on a outcome window of 12-month or more. The *current status* approach was only superior when a 3-month outcome window was used, subject to certain caveats. A drawback to using the *worst status* approach is that it may be slow to recognise recent improvements to customers' financial circumstances. For customers who have previously entered arrears but subsequently recovered, this may result in an unnecessarily prolonged period of default and prevent banks from extending the appropriate support to customers recovering from debt.

Artificial Data

As discussed in Chapter 3, a credit scorecard for retail loan applicants is constructed using a sample of previously accepted loan applicants along with their actual outcomes at some later date. The data describing the loan applicants consists of financial and demographic information. Previously, Section 4.3 highlighted that, for many academic researchers, obtaining real credit scoring data with which to evaluate modelling approaches is a problematic and time-consuming task. It is therefore reasonable to regard the credit scoring research community, like many other research communities, as one with a weak data sharing culture. This is not meant as a criticism of the community itself but rather an acknowledgement of the barriers to sharing data.

In this chapter we aim to address this issue by proposing a framework to generate artificial datasets that can be used in the design and assessment of classification techniques for residential mortgage application credit scoring. This framework can

be used to generate credit scoring datasets of any size that are customised to the requirements of a researcher based on a set of tunable parameters (e.g. good:bad ratio, and custom feature distributions). This in turn could promote: (i) greater participation and diversified perspectives; (ii) replicable experimental findings; and (iii) increased creativity and solution proposals.

To ensure that our framework is sufficiently grounded in reality, datasets are generated using a range of sources:

- Demographic information from the Central Statistics Office, Ireland (CSO, 2010);
- Housing statistics published by the Irish Government Department of the Environment, Heritage and Local Government (DofE, 2008);
- A profile of Irish loan defaulters developed using market data from Moody's Global Credit Research (Moodys, 2010b) and the Central Bank of Ireland (Kelly *et al.*, 2012; Lydon & McCarthy, 2011).

By engaging with a credit scoring expert (see Acknowledgements) and reviewing the relevant literature we select features that are typical of most credit application scorecard models. To generate a dataset, data values for the features are obtained by randomly generating values based on the data distributions specified in the sources above. In order to assign class labels to the generated data a Credit Risk Score is estimated using a set of non-linear, multi-feature rules. The classification complexity is further enhanced by adding random Gaussian noise.

The rest of the chapter is organised as follows. Section 7.1 describes our artificial data generation framework. A demonstration of how the framework can be used

to simulate population drift is provided in Section 7.2. Finally, future work and conclusions are presented in Section 7.3.

7.1 Methodology

Previously, Section 4.3.1 highlighted two approaches to address the lack of availability of real world-data. One such approach is the generation of artificial data without using any real-world data, for which many specialised dataset generators have been described in the literature, e.g. see Alaiz-Rodríguez & Japkowicz (2008); Scott & Wilkins (1999); Srikant (1994). The work in this chapter adopts the same approach of generating artificial data without using any real-world data.

The following section explains the process our framework uses to generate an artificial residential mortgage application credit scoring dataset. A full credit scoring dataset of n instances is produced by first generating the feature values for the n instances and then applying a label to each instance. The process can be decomposed into two stages: (i) Feature Value Generation; and (ii) Label Application, each of which will be explained in detail.

7.1.1 Feature Value Generation

A loan applicant is described by 16 separate features, including both categorical and continuous data types. The features and their attributes were selected based on:

- The advice of an Irish credit risk expert;
- The availability of relevant statistics upon which to base feature value distributions;

- An evaluation of features described in previously published credit scoring literature, e.g. (Anderson, 2007; Leung Kan Hing, 2008);
- A review of mortgage application forms used by Irish banks.

The main sources used to obtain statistics about feature values were: (i) the Central Statistics Office, Ireland (CSO) (CSO, 2010); (ii) the Department of Environment, Heritage and Local Government (DEHLG) (DofE, 2008); and (iii) analysis of securitised Irish residential mortgages by international ratings agencies (Fitch Ratings, 2007; Moodys, 2010a). The 16 features used are listed in Table 7.1 and briefly described below.

Table 7.1: Artificial dataset features

<i>Feature</i>	<i>Type</i>	<i>Description</i>
Location	Categorical	Location of purchased dwelling
New Home	Binary	Newly built dwelling
First-Time-Buyer	Binary	Never purchased property before
Age Group	Categorical	Age of the borrower
Income Group	Categorical	Total income of the borrower
Employment	Categorical	Borrower's employment sector
Occupation	Categorical	Employment activity of the borrower
Household	Categorical	Family composition
Education	Categorical	Highest level of formal education
Expenses-to-Income	Continuous	Ratio of borrower-expenditure-to-income
Loan Value	Categorical	Amount advanced to the borrower
LTV	Categorical	Loan-to-value ratio
Loan Term	Categorical	Length of the loan in years
Loan Rate	Categorical	Interest rate paid on the loan
House Value	Categorical	Market value of the property
MRTI	Continuous	Ratio of mortgage-repayments-to-income

Location This feature describes the location of the property associated with a mortgage application. The 6 attributes used (Dublin, Cork, Galway, Limerick, Waterford and Other) were obtained using data from Fitch Ratings (2007).

New Home This binary feature indicates if the borrower is purchasing a newly built property or a previously occupied property.

First-Time-Buyer (FTB) This binary feature specifies whether or not the borrower has previously purchased property.

Age Group This feature specifies the age of the borrower. The 6 attributes used, based on attributes previously used by the DEHLG (DofE, 2008), are: (i) 18 – 25; (ii) 26 – 30; (iii) 31 – 35; (iv) 36 – 40; (v) 41 – 45; and (vi) 45+.

Income Group The total annual income of the borrower is captured by this feature. Based on data contained in the DEHLG housing statistics (DofE, 2008), 6 attributes are used. These attributes (in '000€) are: (i) 40 – 60; (ii) 60 – 80; (iii) 80 – 100; (iv) 100 – 120; (v) 120 – 150; and (vi) 150+.

Employment This feature represents the employment sector of the primary borrower. This feature is based on data contained in the CSO statistical yearbook (CSO, 2010) whose 14 attributes are derived from the EU NACE Revision 2 classification (Eurostat, 2008).

Occupation This feature attempts to measure the borrower's seniority within their employment sector. The CSO Broad Occupational Groupings (CSO, 2010) and DEHLG housing statistics (DofE, 2008) are used as a basis for the 6 attributes, which are: (i) Manager/Employer; (ii) Office:Salaried; (iii) Skilled; (iv) Semi-Skilled; (v) Manual; and (vi) Self-Employed.

Household The composition of the borrower's household is defined by this feature.

Household is a strong indicator of potential financial outgoings, e.g. childcare fees, university fees. Based on data from the CSO (2010) the feature is split into 6 attributes: (i) 1 Adult, no child < 18; (ii) 1 Adult, 1+ child < 18; (iii) 2 Adults, no child < 18; (iv) 3+ adults, no child < 18; (v) 2 Adults, 1+ child < 18; and (vi) Other.

Education The Education feature captures the highest level of formal education attained by the borrower. The feature is divided into 7 attributes as used in the Irish educational system: (i) Primary or below; (ii) Lower secondary; (iii) Higher secondary; (iv) Post leaving certificate; (v) Third level non-honours degree; (vi) Third level honours degree or above; and (vii) Other.

Expenses-to-Income This feature represents the standard level of borrower expenditure on commodities and services. The data is derived from the most recent CSO Household Budget Survey (CSO, 2010). The attributes used (in '000€) are: (i) 0 – 30; (ii) 30 – 45; (iii) 45 – 60; (iv) 60 – 75; (v) 75 – 90; and (vi) 90+.

Loan Value Group This feature details the principal of the loan. The nine attributes used by Loan Value Group are based on those used by DEHLG housing statistics (DofE, 2008), which are (in '000€): (i) 50 – 100; (ii) 100 – 150; (iii) 150 – 200; (iv) 200 – 250; (v) 250 – 300; (vi) 300 – 350; (vii) 350 – 400; (viii) 400 – 450; (ix) 450 – 900.

Loan-to-Value (LTV) This feature expresses the ratio of the loan value to the market value of the asset. Based on data from DEHLG housing statistics (DofE, 2008) we use 10 attributes: (i) < 45%; (ii) 45% - 55%; (iii) 55% - 60%; (iv) 60% - 65%; (v) 65% - 70%; (vi) 70% - 75%; (vii) 75% - 85%; (viii) 85% - 93%; (ix) 93% - 97.5%; and (x) 97.5% - 100%.

Loan Term The duration of the loan in years is captured by the Loan Term feature. The data is based on DEHLG housing statistics (DofE, 2008). Five attributes are employed: (i) 20-years; (ii) 25-years (iii) 30-years; (iv) 35-years; (v) 40-years.

Loan Rate This feature represents the interest rate paid by the borrower on the loan. For simplicity we do not consider interest-only loans. The attributes used are: (i) Tracker Type 1 - linked to the European Central Bank (ECB) rate for the life of the loan; (ii) Tracker Type 2 - linked to the ECB rate for 10 years or less; (iii) Fixed Type 1 - fixed rate loan for the life of the loan; (iv) Fixed Type 2 - fixed rate loan reverting to standard variable after 5 years or more; (v) Standard Type 1 - standard variable loan rate for the life of the loan; and (vi) Standard Type 2 - fixed rate loan reverting to standard variable after less than 5 years. The data is derived from analysis provided by Moodys (2010a).

House Value This feature represents the market value of the asset. House Value is derived from existing features and is calculated as Loan Value divided by Loan-To-Value. House Value is generated as a continuous value that is then converted in a categorical value. The categorical values are based on attributes used by DEHLG housing statistics (DofE, 2008). The attributes used (in '000€) are: (i) 0 – 150;

(ii) 150 – 200; (iii) 200 – 250; (iv) 250 – 300; (v) 300 – 350; (vi) 350 – 400; (vii) 400 – 500; and (viii) 500+.

Monthly-Repayments-to-Income (MRTI) This is a continuous feature which expresses monthly mortgage repayments as a percentage of monthly income.

The process of generating feature values for an individual instance starts by randomly generating the values for a small number of core features (Location, New Home, and Loan Rate). These are randomly generated based on a set of user-defined prior probabilities, that allow for the incorporation of existing domain information. Similarly, the non-core feature values are randomly generated based on conditional prior probabilities, except for House Value and MRTI which are derived from previously generated feature values. MRTI is derived from existing features as it expresses the ratio of monthly mortgage repayments to monthly income. House Value is calculated as Loan Value divided by Loan-To-Value. For the non-core features, in order to impose realistic assumptions about the generated data values, it is necessary to control the correlation between them. This correlation is implemented through the use of conditional prior probabilities. We emphasise that we do not assume that these correlations capture the full relationship between all of the features, but rather that they provide a useful explanation from which to generate meaningful data.

Table 7.2 describes the conditional prior probabilities of each feature. As House Value and MRTI are both derived from existing features they are omitted. Each feature listed in Table 7.2 is directly correlated with one or more other features.

The reasoning behind each correlation is also provided. The basis for many of these correlations is derived from CSO (2010) and DEHLG (DofE, 2008) publications. In addition, a Moody's report (Moody's, 2010b) provides a further means of validating the conditional prior probabilities. An advantage of our framework is that the user can define both the conditional and prior probability values, allowing for the creation of customised datasets. The conditional prior probabilities and the discussed default settings are detailed in Appendix D. A more in-depth description of the features and their corresponding attributes is available in Kennedy *et al.* (2012a)¹.

To illustrate the process of generating a new instance we will present an example. The first step in generating a new instance is to assign a single Location value randomly drawn from a prior distribution. The default values used are based on an analysis from Fitch Ratings (2007), which specifies the prior probabilities as: Dublin (32%); Cork (15%); Galway (7%); Limerick (4%); Waterford (3%); and Other (39%). Next, a binary flag is randomly applied to indicate whether the loan relates to a New Home or not. The default prior probabilities, based on DEHLG statistics (DofE, 2008), are Yes (46%) and No (54%). For the Loan Rate feature the prior probabilities are specified as per Table 7.3. The remaining feature values are then generated based on the conditional probabilities as described in Section D.2. For example, the distribution of possible values for the FTB feature (Yes and No) is conditional on the Location and New Home features (initially assigned as described above). The conditional prior probability values for FTB are detailed in Table 7.4. Based on this, an instance assigned as a New Home and located in Dublin has a 41% probability of being flagged as a FTB and a 59% probability of being assigned

¹http://www.comp.dit.ie/aigroup/?page_id=729

Table 7.2: The conditional prior probabilities (Dependency) of each feature with a brief explanation

<i>Feature</i>	<i>Dependancy</i>	<i>Explanation</i>
FTB	Location	The high cost of property for certain locations (e.g. Dublin) restricts the number of FTBs
	New Home	On the basis of affordability, a FTB is more likely to purchase a new home
Age Group	FTB	The lower age groups are likely to consist mostly of FTBs
Income Group	Location	Income is determined, in part, by living expenses which vary by location
	FTB	As FTBs tend to be younger than non-FTBs, they are likely to earn less
Employment	Occupation	Occupation strongly influences employment sector
Occupation	Income Group	The occupation of a borrower can be attributed to their level of income
	Location	Senior occupations are more likely to reside in a high density location
	FTB	As FTBs tend to earn less, they are unlikely to be employed in a senior occupation
Household	Income Group	A large income can be attributed to a household with multiple earners
	FTB	A FTB is more likely to be a single person or belong to a young family
Education	Age Group	Younger borrowers tend to have a higher standard of formal education
Expenses-to-Income	Income Group	Expenses are tied to the borrower's income
	Household	The composition of a household affects expenses
Loan Value	Location	Property prices vary by location
	New Home	New homes are likely to be cheaper than existing homes
	FTB	A FTB is likely to borrow less
LTV	Location	Property prices vary by location
	New Home	New homes are likely to be cheaper than existing homes
	FTB	A FTB is unlikely to have as much saved as a non-FTB
Loan Term	Location	Loan size is influenced by regional property prices. Larger loans normally require a longer term
	FTB	FTBs are likely to require a longer loan term
	Age Group	Younger people are likely to earn less and require longer loan terms

as a non-FTB. Similarly, an existing home outside of Dublin has a 30% probability of being flagged as a FTB and a 70% probability of being assigned as a non-FTB. Tables similar to Table 7.4 exist for each other feature and, using these, any number of plausible residential mortgage applicant instances can be generated. Although all prior and conditional prior probabilities are grounded in the the information sources described above, they can all be adjusted to simulate different economic scenarios.

Table 7.3: Loan Rate prior probabilities. The loan rate values used when calculating the monthly loan repayments are also provided.

<i>Loan Rate</i>	<i>Loan Rate Value</i>	<i>Prior Probability</i>
Tracker Type 1	1.50%	15.50%
Tracker Type 2	2.50%	9.45%
Fixed Type 1	5.35%	45.00%
Fixed Type 2	5.00%	6.70%
Standard Type 1	3.50%	14.00%
Standard Type 2	4.50%	9.35%

To further ensure the plausibility of the generated data, a set of configurable rules are used to remove instances that represent scenarios that would never be expected to arise in real life. For example, any instances with an MRTI of 0.8 or more are removed. Further details appear in Section D.3.

7.1.2 Label Application

After the feature values for a set of instances have been generated the next stage is to apply a label (good or bad) to each instance. Figure 7.1 provides an overview of the labelling process. The instances generated following the approach described in Section 7.1.1 act as input to this process.

In Step 1, one instance at a time, a set of *Coded Rules* are used to calculate a *Credit Risk Score* for each instance. The higher the Credit Risk Score, the greater

Table 7.4: A list of conditional prior probabilities (CPP) for the FTB feature. Due to a lack of available statistical information we do not differentiate between the CPP of properties located outside of Dublin, the country’s main population centre.

<i>Location</i>	<i>New Home</i>	<i>FTB</i>	<i>CPP</i>
Dublin	1	1	41%
Dublin	1	0	59%
Dublin	0	1	30%
Dublin	0	0	70%
Not Dublin	1	1	38%
Not Dublin	1	0	62%
Not Dublin	0	1	30%
Not Dublin	0	0	70%

the likelihood of default. A Gaussian noise term (from a user-specified distribution) is then added to the Credit Risk Score in order to simulate uncertainty in the environment (e.g. imperfect data capture). In Step 2 the instances are ordered by their Credit Risk Scores and divided into a user-specified number of equal sized groups. In Step 3, based on a user-specified default rate, a class label for each instance is drawn from a Bernoulli distribution, where the parameters of the distribution are customised for each of the groups created in Step 2. The remainder of this section will explain these steps in detail.

7.1.2.1 Label Application: Step 1

The coded rules used to determine the Credit Risk Score for an instance were developed based on extensive consultations with a credit risk expert, and review of a Moody’s report (Moodys, 2010b) profiling Irish loan defaulters and a Central Bank of Ireland technical report (McCarthy & McQuinn, 2010) on the loan repayment behaviour of Irish mortgage holders. The coded rules take as input the feature values of a single instance and output an overall Credit Risk Score for that instance.

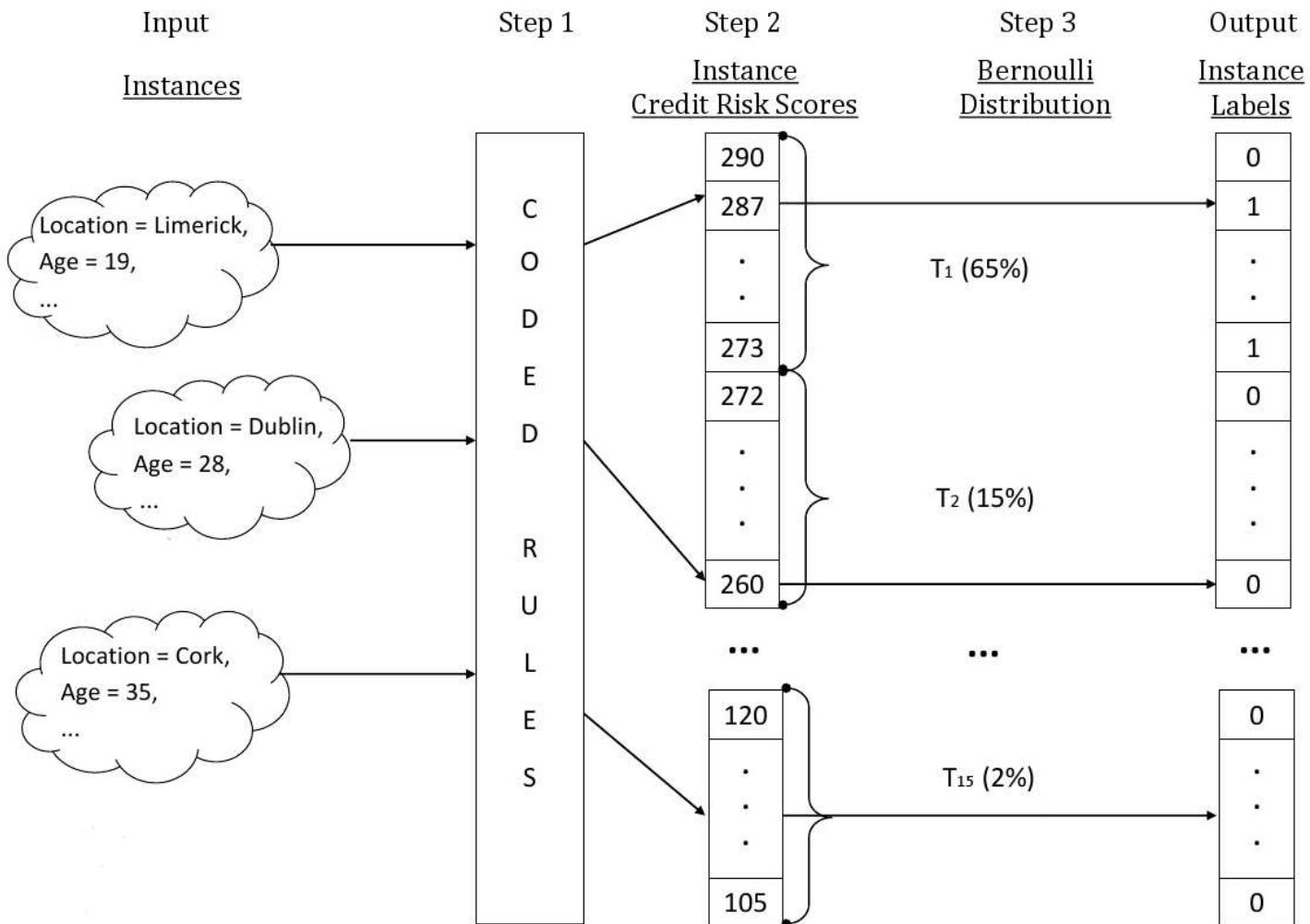


Figure 7.1: Data Labelling process.

For each feature value, the individual coded rules determine the associated *Risk Level* (e.g. Location Risk Level, Age Risk Level), where the risk levels are monotonically ordered in increasing levels of risk (i.e. Level 1 is the lowest level of risk) and are in the range (1, 10). The risk level for a particular feature value is first based on a user-defined look-up table that maps the possible values of a feature to specific risk levels. This initial risk level is then modified based on the values of other features that are deemed to be related to the initial feature. This modification is used in order to make the labels generated for applicant instances more plausible, and to increase the complexity of the labelling process.

To better illustrate this process the application of coded rules to generate a risk level for the Location feature for a mortgage applicant instance with a Location value of *Galway*, a Home Value in the range *400k - 500k* and a *skilled* Occupation is provided in Figure 7.2. In this example, Galway is considered a medium-to-low risk location as it has a low unemployment rate and is considered as a desirable location in which to live. As a result the applicant is given an initial risk level of 4 based on a user-customisable look-up table. The risk level of the Location feature has been defined to interact with the House Value and Occupation features (justification for this is given below). In the context of Location, a high house value (potentially larger amount of negative equity) increases the risk of default. On this assumption, the risk level for the Location feature is defined to rise by 2, up to 6, when the House Value is in the *400k - 500k* range. Finally, the fact that the borrower has a *skilled* (i.e. tradesperson) Occupation is defined to reduce the risk level by 1 as they are more likely to remain in secure employment. This results in a final risk level of 5 for the Location feature. The size of the adjustment to the risk level caused by

House Value and Occupation will differ by Location. Full details of all values used are given in Kennedy *et al.* (2012a).

The rules and specific interactions used to determine the risk levels for each feature used in the framework, which are detailed below, are formed using the knowledge and experience of: (i) a credit scoring expert; and (ii) market data and analysis from Moody's Global Credit Research (Moody's, 2010b) and the Central Bank of Ireland (Kelly *et al.*, 2012; Lydon & McCarthy, 2011; McCarthy & McQuinn, 2010). The features with which the target feature has an interaction are listed in parentheses.

i. **Location (House Value, Occupation):** The ability to sell or rent a house can reduce the likelihood of default. Dublin and Cork are the main rental markets in Ireland and as such represent a lower risk of default. House Value is used to indicate that for some locations, due to demand, houses may be prone to a higher valuation (e.g. Dublin in 2007). The more overvalued a home, the greater the risk of default on account of the negative equity that may arise as house prices return to their long-term average (Moody's, 2010b).

Occupation, in the context of Location, considers that the level of demand for a borrower's expertise and experience varies from location to location. Typically, a populous location indicates a large and diverse jobs market.

ii. **MRTI (Loan Rate, Expenses-to-Income):** In general, when granting credit, a borrower's MRTI should be no more than 31% (see Kelly *et al.*, 2012). The Loan Rate is one of the main variables used when calculating the monthly mortgage repayment amount. The Expenses-to-Income ratio and

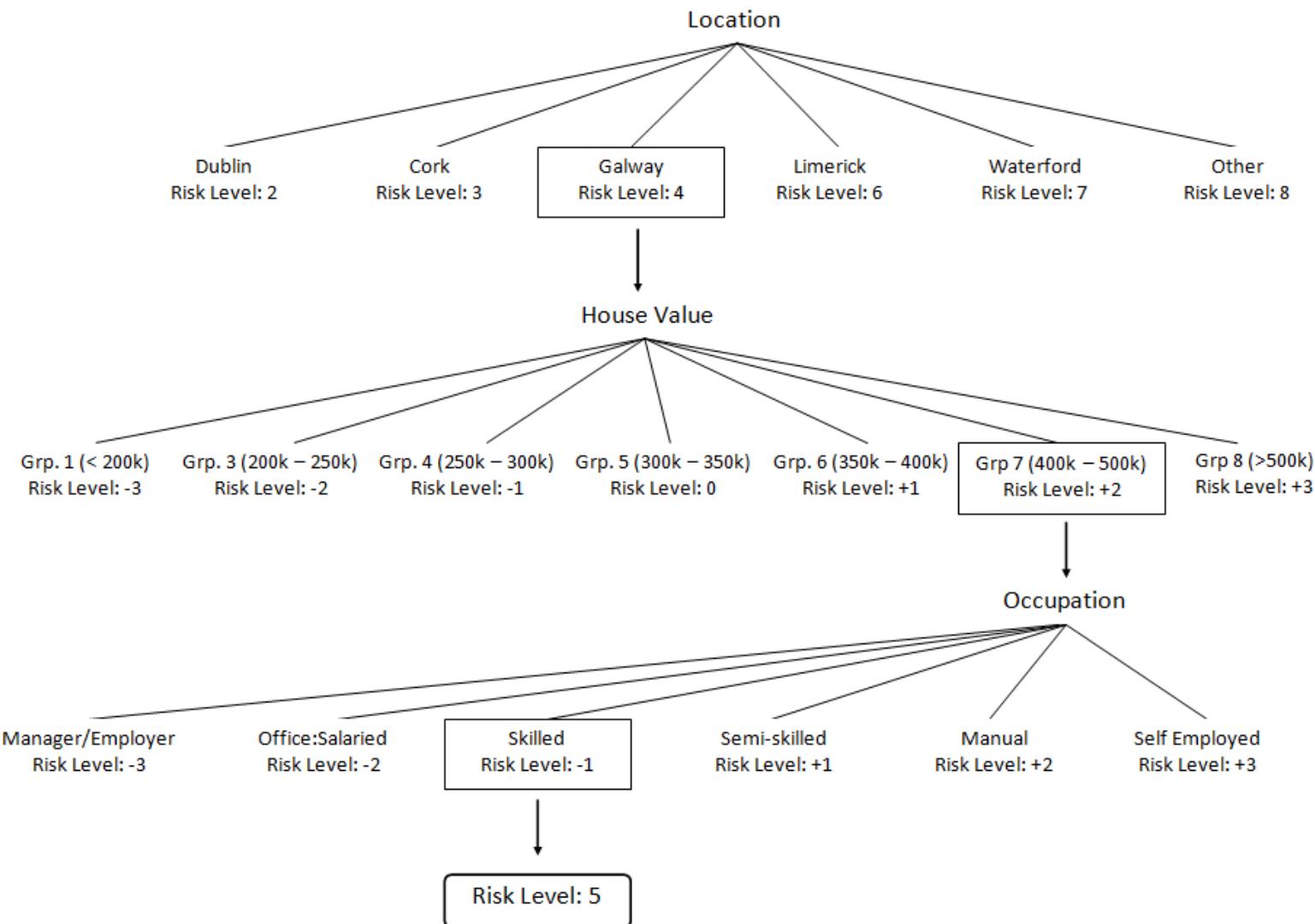


Figure 7.2: Example of Coded Rules: Calculation of the risk level for the Location feature for a mortgage applicant instance with a Location value of *Galway*, a Home Value in the range *400k – 500k* and a *skilled* Occupation.

MRTI provide a clear indication of a borrower's overall expenditure. A borrower with a relatively high MRTI (33% - 38%) may initially appear as a risky prospect. However, this risk is reduced somewhat if their Expenses-to-Income ratio is low and their Loan Rate is fixed (i.e. the loan repayments are not subject to any variability for the foreseeable future).

iii. **Loan Value Group (Income Group, LTV):** The rules are coded under the assumption that the higher the loan value the greater the risk of default, as per Moody's report (Moody's, 2010b). This risk, however, may be offset by a high level of Income or a low LTV. A high level of income suggests that the borrower has the financial means to service a large loan. A low LTV reduces the risk of default, as the borrower has already invested too much in the loan to simply walk away.

iv. **Employment (Location, Education)** A borrower's Employment sector represents, to some degree, their job security and earnings. Location is deemed to impact Employment risk as it relates to the availability of commensurate employment opportunities within the same locale, i.e. the size of the jobs market for a particular Employment sector. For example, the hospitality industry of Galway is larger than that of Limerick or Waterford, and as a consequence is considered less of a risk in terms of default. Education may indicate the mobility of the borrower in terms of finding employment in alternative Employment sectors. A higher level of formal education should increase the likelihood of mobility between Employment sectors. High mobility is perceived as reducing the likelihood of default.

v. **Occupation (Employment, Expenses-to-Income)** The demand for a particular Occupation varies for each Employment sector. Based on Moodys (2010b) we assign borrowers assigned to the *Self-employed* Occupation category as the most likely to default. Borrowers belonging to the *Manager/Employer* Occupation category are considered the least likely to default - due to their importance to an organisation as well as the demand for their skills and experience. By using a borrower's Occupation along with their Expenses-To-Income the coded rules attempt to capture the borrower's social status and the cost of maintaining it.

vi. **Income Group (Household, MRTI)** A borrower with a high level of income is considered less likely to default. Household points to the level of income required to maintain the household (i.e. the number of dependants). The MRTI indicates the amount of income required to service the loan. Together, Household and MRTI, indicate the room for adjustment available to the borrower to changes in financial circumstances and so are used to modify Income Group risk.

vii. **Household (Age Group, MRTI)** Household affects the risk of default with regard to the earning power and priorities of household members. A *Single Person* household represents a greater risk of default compared to the *2 Adults, No Child* category as the impact of a loss of income to the couple may be less severe. From a Household perspective, the Age Group represents the level of responsibilities the borrower may have. For example, a middle-aged household with dependents may have to use their savings to pay for tuition fees, thereby

increasing the risk of default. The MRTI indicates the burden the household is under to pay bills.

viii. **Age Group (Income Group, LTV)** Younger borrowers represent a greater risk of default than older borrowers. Income Group, in the context of Age Group, indicates the present and future earning potential of a borrower. A young person on a high income can be interpreted as a skilled individual and therefore low risk. As the LTV reflects the amount of savings a borrower has contributed to the loan, the Age Group indicates how long the borrower has had to accumulate wealth. An older borrower with a low LTV can be interpreted as one with access to savings and therefore lower risk.

ix. **FTB (Loan Value Group, Loan Rate)** Due to a lack of experience, a First-Time-Buyer is considered more likely to default on a loan than a non-FTB. A high Loan Value Group places greater pressure on the borrower to manage their finances, thereby increasing the risk of default. The Loan Rate indicates the borrower's financial aptitude at selecting an appropriate loan product. For example a FTB with a fixed rate is considered risk averse compared to a FTB with a variable rate.

x. **Education (Income Group, Occupation)** A borrower's level of formal Education impacts the risk of default. The more educated an individual, the more financially astute they are likely to be. A high level of education also improves job prospects. Income Group affects Education in terms of the ability to afford additional training and improve skill levels. A borrower's Occupation indicates their ability to successfully apply their education, i.e. over-achiever

or under-achiever.

xi. **Loan Rate (House Value, Loan Value Group)** There are three different Loan Rates: (i) Fixed rate loans which are considered the least risky; (ii) Tracker rate loans which are slightly more risky; and (iii) Variable rate loans which are considered the riskiest of the three. In terms of the Loan Rate, the Loan Value Group affects risk based on the fact that the size of the repayments increase with the size of the Loan Value. In general, interest rates affect the availability of capital and demand for investment. If interest rates rise an expensive house will become harder to sell, thus rising interest rates may cause the house value to decline, thereby increasing the possibility of negative equity.

xii. **Expenses-To-Income (Household, Age Group)** The greater the Expenses-to-Income ratio the higher the risk of default. Household is a strong indicator of how much money the borrower needs to spend on groceries, utility bills, fees etc. The Age Group helps identify the level and the necessity of the expenses.

xiii. **LTV (FTB, Occupation)** When assessing LTV for the risk of default the coded rules consider the size of the deposit provided by the borrower. In the event of a default, a high LTV indicates the borrower will suffer less of a loss compared to someone who has already invested a large amount of savings. A non-FTB with a high LTV may indicate poor financial management, or over-stretching when trading up. Based on reputation, higher ranked Occupations are able to receive a higher LTV and should not be penalised as such, thus preventing a rise in the risk of default.

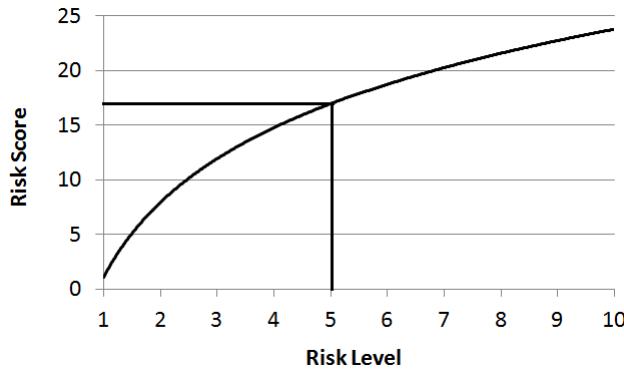


Figure 7.3: Curvilinear transformation: The Location risk level (5) detailed in Figure 7.2 is converted into a Location risk score (17.05) using a user-definable transformation function.

xiv. **House Value (LTV, MRTI)** The greater the house value the less likely the borrower will default. Other factors to consider when assessing risk based on the house value include the deposit paid by the borrower and the amount of income required to service the loan.

After risk levels are calculated for each feature they are each transformed using a user-customisable transformation function to generate a corresponding *Risk Score*. Converting risk levels into risk scores is used to add non-linearity to the resulting classification problem, again making it closer to what is likely to be encountered in real credit scoring scenarios. The default transformation function has a monotonically increasing shape, this ensures that with each level the risk score gradually increases. Figure 7.3 shows an example where the risk level of 5 for the Location feature detailed in Figure 7.2 is transformed, using a curvilinear transformation function, to a risk score of 17.05.

These individual feature risk scores are then summed to generate the overall Credit Risk Score for an instance. Finally, a random Gaussian noise term, drawn

from a user defined distribution¹, is added to the risk score to simulate the noise inherent in real world risk scenarios. Table 7.5 illustrates the way in which the Credit Risk Score of an instance is calculated based on the individual risk scores along with the addition of a noise value. In this study we assign equal weighting to each feature risk score, this implies that all features are of the same importance. Through user-defined parameters it is possible to adjust the importance of an individual feature. To help ensure that the generated data is realistic, instances with an MRTI greater than 0.8, or an affordability (calculated as outstanding loan amount divided by annual income) greater than or equal to 11 are removed. This typically accounts for 2% to 2.7% of the data, however this value can vary as both settings are user-adjustable.

Table 7.5: Credit Risk Score of a single instance

<i>Feature</i>	<i>Feature Value</i>	<i>Risk Level</i>	<i>Risk Score</i>
Location	Dublin	3	11.94
MRTI	27.40%	6	18.87
Loan Value	€385,000	8	21.75
Employment	Retail	8	21.75
Occupation	Self-Employed	9	22.93
Income	€102,000	5	17.05
Household	2 Adults, no child < 18	5	17.05
Age	47 years-old	8	21.75
FTB	No	6	18.87
Education	Lower secondary	8	21.75
Loan Rate	5%	8	21.75
Expenses-to-Income	47%	5	17.05
LTV	70%	7	20.41
House Value	€550,000	2	7.88
Risk Score Total			260.78
Noise			-0.23
Credit Risk Score			260.55

¹as a default we use a mean of 0 and a standard deviation of 0.25

7.1.2.2 Label Application: Step 2

In Step 2 of the labelling process, once Credit Risk Scores have been calculated for every instance, the instances are sorted in descending order of their Credit Risk Scores and divided into a maximum of 15 equally sized groups (a smaller number of groups may be used). This number was selected based on numerical experimentation showing no significant change in the classification accuracy of a logistic regression classifier given an artificial dataset generated using the default settings.

7.1.2.3 Label Application: Step 3

In Step 3 an overall default rate is specified (e.g. 2.8% reflects the Irish mortgage market in 2008). Each group (T_1 to T_n) is assigned a user-defined default rate (μ_i), which specifies the proportion of the overall default rate to come from that group. As T_1 consists of instances with the highest Credit Risk Score it is assumed to contain the largest default rate (i.e. μ_1). For T_1 to T_n , each instance is assigned a random label drawn from a Bernoulli distribution with a proportion μ_i of the outcomes equal to 1 (bad) and the remaining proportion $1 - \mu_i$ with outcomes equal to 0 (good). The element of randomness used by the labelling process attempts to simulate unforeseen circumstances (e.g. divorce, death, or a personal financial shock) that a borrower may experience. It should be noted that adjacent groups with the same user-defined default rate are effectively merged.

For example, consider the process of labelling a dataset of 3,000 instances with an overall default rate of 2.8% (84 instances or 2.8% of 3,000), as shown in Figure 7.1. The instances are sorted into 15 equally sized groups (200 instances per group)

based on their Credit Risk Score. Based on a user-defined setting, 65% of the defaulters (55 instances) are located in T_1 . The instances in T_1 are then labelled as good or bad based on a Bernoulli distribution with a probability of being bad equal to 0.275 (55/200). Similarly, based on a user-defined setting, 15% of the defaulters (12 instances) are located in T_2 , hence a Bernoulli distribution with a probability equal to 0.06 (12/200) is used to label the instances as bad. The same approach is used to label the remaining 20% (17 instances) of defaulters distributed between T_3 and T_{15} . The labelling process described provides the experimenter with the means to systematically manipulate the structure of the artificially generated dataset.

7.1.3 Summary

This section has described the methodology used to generate artificial mortgage application credit scoring data, and the process used to label this data. This process has been designed to generate plausible datasets that result in classification problems of similar complexity to those seen in real data scenarios. Using our methodology instances are described by a collection of feature values (Location, New Home etc.) that are based on those most commonly used in real credit scoring problems. A Credit Risk Score based on the values of these features (where higher scores imply greater risk) is calculated for each instance by the application of a collection of coded rules that have been devised through consultations with a credit risk expert and a review of relevant literature. Ordering the instances by their Credit Risk Scores into groups, each with a specified default rate, instances are randomly labelled as good or bad which allows for the simulation of factors not included in the model, and further complicates the resulting classification problem. At every point in the development

of this process we have sought to base decisions on as much credit scoring data as possible.

As both the feature value generation and label application processes are heavily user-customisable, it is possible to create any number of datasets representing a whole range of credit scoring scenarios. This allows researchers to efficiently construct credit scoring datasets on which a wide range of experimentation can be performed. Section 7.2 describes an illustrative example in which datasets are generated to explore the problem of population drift in credit scoring.

7.2 Illustrative Example: Population Drift

This section describes the application of our artificial data generation framework to the study of population drift in mortgage application credit scoring. Credit scorecards have a limited lifespan, and often their performance degrades over time. During credit scorecard construction samples drawn from data representative of the current population will rarely have the same distribution as those drawn from future populations. When one data source, S_1 , changes to another, S_2 , *population drift* is said to have occurred. Changes in the underlying population pose a serious problem in practical fields such as finance, medical diagnosis and bioinformatics (Hand, 2006b). In the credit scoring domain, because of its competitive environment, this problem is particularly acute (Hand, 2006b). If left undetected, population drift can result in costly, inadvertent, and unforeseen effects as misinformed strategic decisions are made based on inaccurate tools.

Population drift in credit scoring is a difficult area to study, firstly because it is

difficult for researchers to persuade financial institutions to share confidential and sensitive commercial data that spans a number of years, and secondly because the ways in which populations change in real data are rarely well understood which makes it difficult to draw conclusions about how best drift should be handled. As a result only a limited number of studies addressing the area have been performed (see Kelly *et al.*, 1999; Pavlidis *et al.*, 2012) and the area is ripe for the application of artificial datasets.

In this section we describe how our framework can be used to generate a number of artificial datasets where the distribution of the features gradually changes over time and so controlled population drift is displayed. The impact of population drift is seen in a degradation of classifier performance as the underlying feature distributions change. We make no effort to correct the problem, but rather simply show how datasets useful for studying population drift can be easily generated using our framework.

7.2.1 Framework Configuration

The aim of this demonstration is to show that our framework can be used to generate a dataset exhibiting controlled population drift that can be used to test drift handling or drift identification approaches. The presence and impact of population drift is illustrated by the performance of a logistic regression (LR) model trained on (i) a batch of artificially generated data in which the generation process has been adjusted in order to represent a changing scenario and (ii) a batch of artificially generated data in which the generation process remains static, and so does not exhibit any population drift. Logistic regression is selected as it is commonly used to build

credit scoring models (Bellotti & Crook, 2009). The LR model was implemented using the Weka (version 3.7.1) machine learning framework (Witten & Frank, 2005).

In generating all of the datasets described in this section, default values for all prior and conditional prior probabilities were used (see Kennedy *et al.*, 2012a). In the label application phase of data generation an overall default rate of 1.85% was specified. This figure is representative of the Irish mortgage market in 2007, upon which the initial values for the conditional and prior probabilities used in the feature value generation phase are based. Gaussian noise terms, obtained using the default settings, are added to the Credit Risk Scores of each instance.

Initially an artificial dataset (3,570 instances) was generated and divided into two subsets: (i) the training set (70%); and (ii) the validation set (30%). To limit the scope of the study we do not perform feature selection (see Section 3.2.2.1) or coarse classification (see Section 3.2.2.2) on the data. The training set and the validation set were used to train and tune the logistic regression classifier. Tuning involved optimising the logistic regression ridge estimator parameter in order to offset unstable coefficient estimates that arise from highly correlated data (Le Cessie & Van Houwelingen, 1992). Next, in order to create a control benchmark, 15 additional datasets (2,500 instances each) were generated using the same conditional and prior probabilities and parameter settings as those used in generating the training data. As no population drift was simulated, we refer to these datasets as the *Non-drift* datasets. The prediction performance of the LR model on each of the 15 non-drift datasets was measured using the area under the receiver operator curve (AUC) (see Bradley, 1997).

To simulate population drift, the conditional and prior probabilities of the fol-

lowing features are adjusted to generate a second series of test datasets: (i) Location; (ii) New Home; (iii) FTB; (iv) Age Group; (v) Occupation; (vi) Employment; (vii) Education; (viii) Expenses-to-Income. These features are selected as they capture the demographic information of the population. The adjustments occur over 5 phases with 3 datasets (each containing 2,500 instances) generated per phase. These datasets are referred to as the *Drift* datasets.

For the first phase, the conditional and prior probabilities are the same as those used to generate the training data. In the subsequent phases the user-defined values for the conditional and prior probabilities are altered. For each feature, the size of this alteration varies from category to category. As an example, Figure 7.4 to Figure 7.6 illustrate the change in the distributions of the Location, FTB, and Age Group features for each of the 5 phases. The distributions of the other 5 features change in a similar manner. The rate of change to the feature distributions in each phase is uniform. Changes to the distributions are a result of adjustments to the user-defined conditional and prior probabilities. Figure 7.7 plots the probability density function of the Credit Risk Scores for Phase 1, Phase 3, and Phase 5. These show that the adjustment to the conditional and prior probabilities of the selected features causes a steady increase in the Credit Risk Scores over the 5 phases. This indicates that the risk profile of borrowers is increasing over time, and suggests that the trained model may need to be adjusted accordingly.

Again, the performance of the LR model on each of the 15 drift datasets was measured. Each test was conducted 10 times using different randomly selected training and validation set splits and the results reported are averages of these 10 runs.

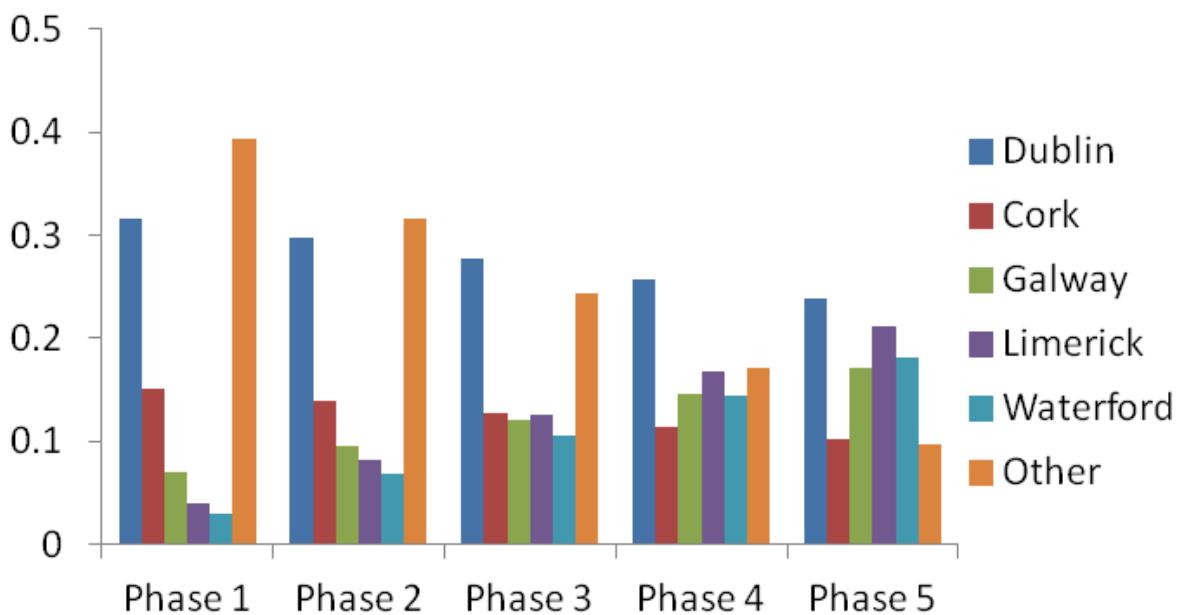


Figure 7.4: Histogram of Location feature for drift datasets. By way of adjusting the prior probabilities of the Location feature, this scenario simulates gradual decrements to the Dublin, Cork, and Other locations along with gradual increments to the Galway, Limerick, and Waterford locations.

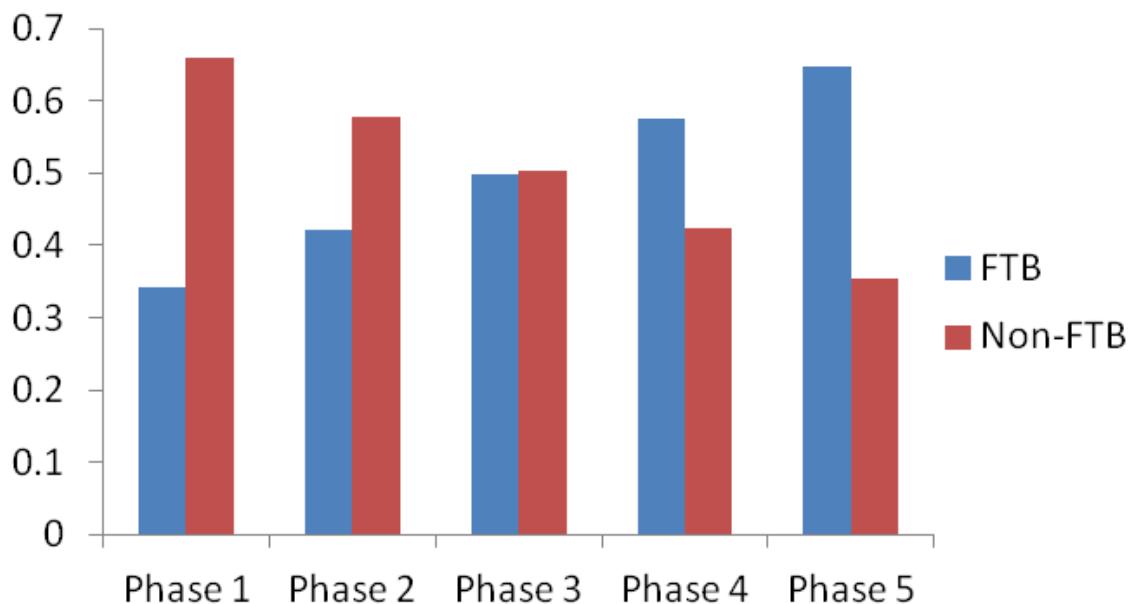


Figure 7.5: Histogram of FTB feature for drift datasets. The prior probabilities of FTB and non-FTB are gradually reversed.

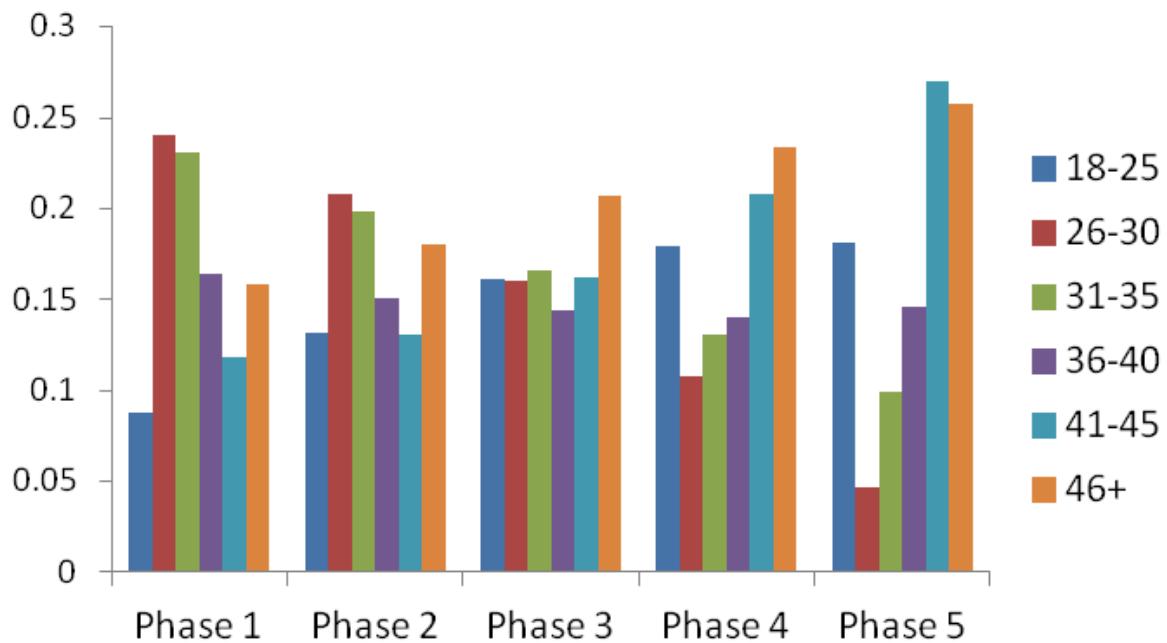


Figure 7.6: Histogram of Age Group feature for drift datasets. The conditional prior probabilities of the two attributes representing 26-to-35 year-olds (26-30, 31-35) are reduced in each phase. The conditional prior probability 36-40 Age Group category remains unchanged. The remaining Age Group attributes (18-25, 41-45, 46+) are increased in each of the 5 phases.

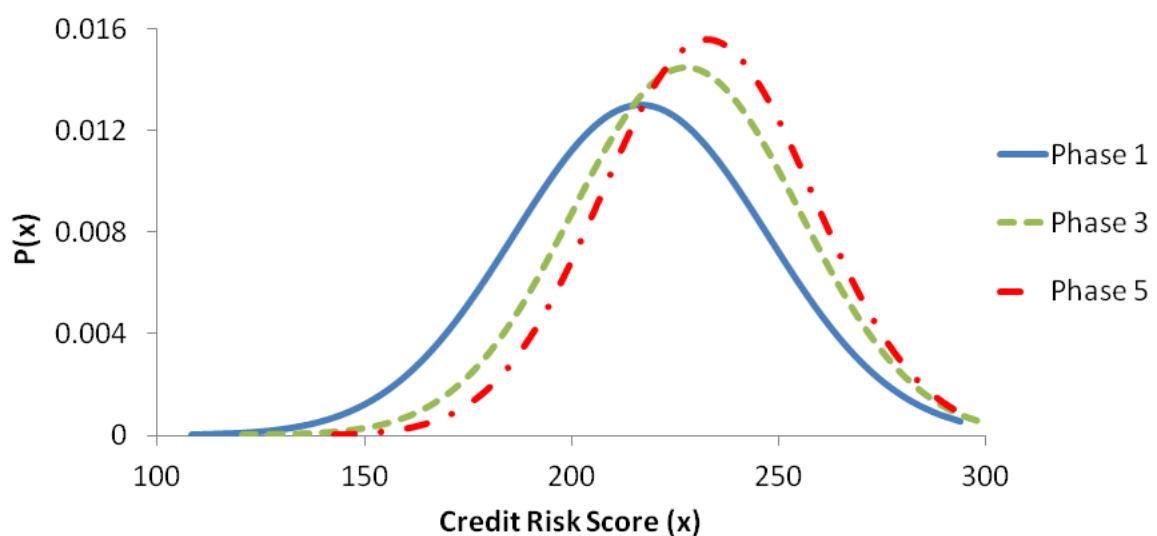


Figure 7.7: Probability density function (PDF) of Credit Risk Score in Phase 1, Phase 3, and Phase 5

7.2.2 Outcome

In this study we use the AUC to measure the performance of the LR model. The AUC is commonly used in credit scoring to estimate the performance of classification algorithms in the absence of information on the cost of different error types. Figure 7.8 shows the performance of the LR model on the non-drift and drift datasets as measured by the AUC (y -axis) over the five phases (x -axis).

The results displayed in Figure 7.8 indicate that the performance of LR model using the non-drift datasets remains constant over the 5 phases. The performance of the LR model using the drift datasets remains robust to the first set of adjustments made to the conditional and prior probabilities (i.e. phase 2). However, as the population drift gradually increases the performance of the LR model steadily declines. This is because the distribution of the samples in the drift datasets have moved significantly away from those used to train the credit model - i.e. population drift has occurred.

This series of datasets can be used to examine both population drift detection mechanisms and different approaches to handling population drift. The use of artificial datasets (as long as they are broadly representative of real datasets in the same domain) for this kind of investigation has two significant advantages over real datasets. First, the occurrence of drift can be pinpointed which is beneficial for evaluating drift detection approaches and is almost impossible with real datasets. Second, the amount of drift exhibited in the data can be controlled by adjusting the data generation parameters and so techniques can be evaluated in the presence of population drift of varying degrees.

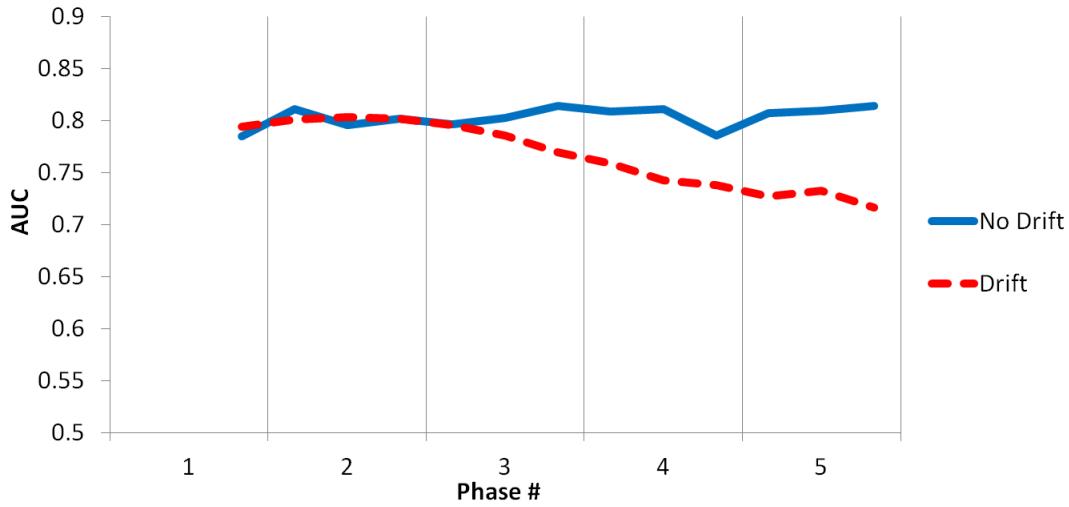


Figure 7.8: The performance of the logistic regression model on the drift and non-drift datasets generated. Performance is measured using the AUC based on a moving average over 3 datasets (where each Phase consists of 3 datasets).

7.3 Conclusions

In this chapter we have described an artificial data generation framework for mortgage application credit scoring scenarios. The characteristics of the data generation framework are based on reputable information sources and the ability for users to adjust parameters in the framework allows the generation of realistic datasets with which to assess the performance of classification techniques.

We make no claim that artificial data can be expected to replicate the rich structural complexities of real-world data. As such, one must exercise caution when using artificial data to assess the superiority of one classification method over another. Indeed, Japkowicz & Shah (2011) advise against such assessments. Artificial data can, however, assist researchers in many ways. For example, we demonstrate how the artificial data framework can be used to show the effects of population drift on the performance of a logistic regression model. The primary advantage artificial data

holds over real-world data is the freedom to design experiments under which certain conditions and parameters can be accurately controlled. Ultimately, the aim of our framework is to assist credit risk researchers in the design of informed experiments with which to test specific hypotheses.

The population drift example is just one illustration of the uses that a parametrised data generation framework for credit scoring data can be put to, and is included to demonstrate the value of such a framework. For future research the example could be extended by examining the effects of feature selection, coarse classification and techniques to handle missing data. Other examples involve experimental studies that use artificial data to compare the accuracy and relevancy of various performance measures and exploitation in a classroom environment to demonstrate binning techniques, feature selection, and sampling techniques. As previously stated, it must be stressed that any conclusion found using artificial data must be verified using real data.

CHAPTER 8

Conclusions

8.1 Introduction

In this thesis we examined three challenges encountered by both practitioners and researchers during the development of quantitative credit scorecards. First, classifier performance was examined with respect to the low-default portfolio (LDP) problem. This also involved assessing the suitability of a number of approaches used to address the LDP problem. Next, we quantified the differences in classifier performance arising from various implementations of a real-world behavioural scoring dataset. The variations were achieved by adjusting the duration of customer behaviour, the outcome period used to label a borrower, and the label definition approach. Finally, we described in detail a framework used to generate artificial data suitable for application credit scoring.

In this final chapter, we briefly summarise key contributions of this work and

discuss several future work directions.

8.2 Summary of Contributions and Achievements

In this thesis, based on detailed results and analyses, we have presented a number of arguments as to why this work is an important contribution to the credit scoring community. The purpose of this section is to gather these arguments, at the end of the thesis, in order to provide the reader with a clear and precise sense of how this work contributes to credit scoring research. The contributions of this thesis include:

- *The identification of the best classifier to use for imbalanced credit scoring datasets* - Chapter 5 described a benchmark study of the performance of supervised classification techniques on a collection of imbalanced credit scoring datasets. Typically, supervised classification algorithms assume a balanced distribution of the classes and attempt to maximise the overall classification accuracy by predicting the most common class. For imbalanced datasets, this assumption can adversely affect the performance of most classification algorithms. To demonstrate the effects of class imbalance in credit scoring we compared the performance of various supervised classification techniques over a range of class imbalances. Comparative experiments showed that the logistic regression classifier performed best.
- *An evaluation of the effectiveness of two commonly used methods for addressing class imbalance* - In Chapter 5 we determine what improvement, if any, can be obtained by: (i) oversampling the minority class; and (ii) adjusting the threshold on classifier output. Oversampling is one of the most popular

solutions to the class imbalance problem. By contrast, many studies avoid the problem of choosing a specific classification threshold by using the AUC, which compares classifier performance over a range of thresholds. In practise this cut-off decision is dependent on a number of factors, including: (i) what aspect of classifier performance is being examined; (ii) the relative cost ratio of false positives and false negatives; and (iii) the strategic considerations of the bank. Regardless, using the AUC may be a problem when we are interested in classifier performance over a narrow range of classification thresholds. Experimental results on various imbalanced credit scoring datasets showed that, for the best performing supervised classification algorithms, oversampling produces no overall improvement. In contrast, adjusting the threshold value on classifier output yields, in many cases, an improvement in classification performance.

- *The use of one-class classification techniques as a solution to the LDP problem*
 - Chapter 5 detailed a benchmark study of the performance of semi-supervised classification techniques on a collection of credit scoring datasets which had been modified to replicate the low-default portfolio problem. In credit scoring, a particularly severe form of class imbalance is referred to as the low-default portfolio problem. To address this issue, we compared the performance of a number of one-class classification (OCC) algorithms with that of logistic regression. In the absence of any statistically significant differences between the results of the OCC techniques and those of logistic regression, we contend that both approaches merit consideration when dealing with LDPs.

- *The identification as to what duration of customer behaviour to use in behavioural scoring models* - Chapter 6 examined to what extent the use of different performance window sizes for model training impacts the classification performance of a behavioural scoring model. Commercial sensitivities surrounding the use of behavioural scoring data ensure that there are very few published empirical studies which directly address this issue. The performance of three separate logistic regression models, each one trained with data based on a particular performance window size, were compared using each of the five possible outcome window sizes. We consider a logistic regression model trained with data based on a 12-month performance window as best suited to the classification task - particularly when outcome window sizes of 3-months, 6-months, and 12-months were specified.

- *The quantitative comparison of classifier performance using different sized outcome periods* - In Chapter 6 the length of the outcome period, from which a borrower's class label is defined, was varied in order to quantify differences between the performance results of 5 separate logistic regression models. Typically, specific business requirements determine the size of the outcome window. However, guidance in the literature on the impact of choices made for this key behavioural scoring modelling parameter is largely limited to anecdotal evidence that suggest good practices and processes for implementation. Our work provides practitioners with a source of comparison for behavioural scoring. Experiment results suggest that the real-world credit bureau data does not contain information which allows a logistic regression classifier to

reliably distinguish between loan repayers and defaulters beyond a 6-month outcome window.

- *The differences between alternative approaches used to define a customer's default status in behavioural scoring* - Chapter 6 compared the classification performance of behavioural scoring models trained using either the *current status* or *worst status* label definition approach. In practice, a lender may often select a particular label definition approach due to some business objective, or to overcome a shortage of history on arrears, or because of regulatory requirements. The purpose of this comparison is to provide practitioners with a published benchmark. Results indicated that a logistic regression model using the *worst status* approach achieved greater accuracy using longer outcome windows than with the *current status* approach.

- *Address the lack of data sharing in credit scoring by describing a framework to generate artificial data for application credit scoring* - Chapter 7 described a framework capable of generating artificial datasets that can be used in the design and assessment of classification techniques for residential mortgage application credit scoring. Due to privacy and commercial sensitivities, the credit scoring research community, like many other research communities, is regarded by some as one with a weak data sharing culture. We demonstrate how the artificial data generation framework allows researchers to design experiments under which certain conditions and parameters can be accurately controlled.

8.3 Open Problems and Future Work

The research presented in this thesis has contributed to the fields of machine learning and credit scoring. However, as with any form of research, each contribution naturally generates more questions. This section introduces some of the outstanding issues for further research which are closely related to this thesis.

8.3.1 Low-Default Portfolios

It is important to note that the two-class classification methods are based on modelling both the distribution of past loan repayers and past defaulters. Whereas one-class classification methods are modelled solely on the distribution of past loan repayers. In situations of *population drift*, where the behaviour of defaulters changes over time due to unrecorded macro-economic factors or, indeed, personal reasons, then the performance of the two-class classifiers will deteriorate. This has been proven by Juszczak *et al.* (2008) in the field of fraud detection whose findings indicate that supervised classifiers, to some degree, over-fit the current training dataset such that when drift is introduced to the class distributions, the supervised classifiers deteriorate faster than the semi-supervised classifiers. This has serious implications for areas such as microcredit. Consider payday loans which are typically small, short duration (less than one month) with extremely high interest rates. It is necessary to construct scorecards that can respond in a timely fashion to shifts in economic and market behaviour, as well as to sudden changes in the borrower's circumstances and behaviour (Thomas, 2009b). Clearly one-class classification is suited to such tasks.

Future work should concentrate on situations for which OCC is well suited.

OCC is best applied in situations with a heterogeneous non-target class where it can be difficult to model or obtain representative training examples. In retail loans the reasons for defaulting are typically unvarying across the portfolios (e.g. loss of income, loss of job, marriage breakdown, poor health). However for models which include economic and market conditions and can thus experience differing scenarios of an economic cycle, two-class classifiers may not be able to model all heterogeneous loan defaulters. OCC approaches might therefore be better suited to these latter problems. Future work could also look at more sophisticated OCC techniques that can utilise small amounts of non-target data (e.g. see Ghasemi *et al.* (2011)).

A more sophisticated form of oversampling, such as SMOTE (Chawla *et al.*, 2002), could also be examined. Another feature of oversampling to consider is the class distribution ratio. Khoshgoftaar *et al.* (2007) reported that an even distribution is not always optimal when dealing with data rarity. To ensure a more representative minority class, clusters could be identified in the minority class from which to sample the data.

8.3.2 Behavioural Scoring

Future work should concentrate on comparing the performance of classification models with duration models such as survival analysis. The work could also be expanded to identify customer accounts that settle early using survival analysis techniques. Furthermore, future work should assess the suitability of credit bureau data as a fundamental risk driver capable of determining defaults. Finally, an additional label definition approach which defines a bad based on a certain percentage of the arrears amount and the outstanding loan value should also be investigated. A topic worthy

of further attention is the ability to identify features used to build classifiers that can determine whether or not customers already in default will recover to repay their loan.

8.3.3 Artificial Data

We made several assumptions in the coding of our rules, for future work some of these assumptions may be reconsidered. Analysis of real-world data should be used to further refine the artificial data in terms of the variability of the data generated, particularly in terms of generating large-scale datasets.

Other potential uses of the framework include generating data to examine problems associated with imbalanced data. In this scenario, artificial data can assist in determining the robustness and sensitivity of a credit scorecard model.

APPENDIX A

Notation

x, x_i : an example

X : set of all examples

y : an output value

Y : set of all output values

S : training set of examples and output values

h : a mapping function $h : X \rightarrow Y$

H : set of all mapping functions

\mathbb{R} : set of real numbers

C : set of classes

c_i : class i

w : weight vector

w^T : transpose of weight vector

p_g : probability of belonging to the *good* class

\exp : exponential function

b_i : regression coefficient

θ : threshold on a probability or a distance

z, z_i : an example or object

μ : mean vector of a dataset

Σ : covariance matrix of a dataset

APPENDIX B

Abbreviations

AE	Auto-encoder one-class classifier
AI	Artificial Intelligence
AUC	Area Under the receiver operating characteristic Curve
BCBS	Basel Committee on Banking Supervision
CFS	Correlation-based Feature Selection
CSO	Central Statistics Office
DEHLG	Department of Environment, Heritage and Local Government
EAD	Exposure At Default
ECA	Expert Committee Approval
ECOA	Equal Credit Opportunity Act
EL	Expected Loss
EM	Expectation-Maximisation

EU NACE	European Union Nomenclature of Economic Activities
FTB	First-Time-Buyer
Gauss	Gaussian one-class classifier
ICB	Irish Credit Bureau
i.i.d.	independent and identically drawn
IRB	Internal Ratings Based
IT	Information Technology
IV	Information Value
k -NN	k -Nearest Neighbour
LDA	Fisher's Linear Discriminant Analysis
LDC	Linear Bayes Normal
LGD	Loss Given Default
Lin SVM	Linear Support Vector Machine
LOG	Logistic Regression
LR	Logistic Regression
LTV	Loan-To-Value
MLE	Maximum Likelihood Estimation
MOG	Mixture of Gaussian one-class classifier
MRTI	Monthly-Repayments-To-Income
NB	naïve Bayes
NN	Neural Network
NParzen	naïve Parzen one-class classifier
OCC	One-Class Classification

PD	Probability of Default
QDC	Quadratic Bayes Normal
ROC	Receiver Operating characteristic
RWA	Risk Weighted Assets
SVDD	Support Vector Data Description one-class classifier
SVM	Support Vector Machine
UCI	University of California Irvine
UK	United Kingdom
USA	United States of America
WoE	Weight of Evidence

C

APPENDIX

Additional Material for Chapter 5

The datasets used in Chapter 5 were obtained from the following sources:

- i. Australia - UCI dataset accessed at [http://archive.ics.uci.edu/ml/datasets/Statlog+\(Australian+Credit+Approval\)](http://archive.ics.uci.edu/ml/datasets/Statlog+(Australian+Credit+Approval))
- ii. German - UCI dataset accessed at [http://archive.ics.uci.edu/ml/datasets/Statlog+\(German+Credit+Data\)](http://archive.ics.uci.edu/ml/datasets/Statlog+(German+Credit+Data))
- iii. Japan - UCI dataset accessed at <http://archive.ics.uci.edu/ml/datasets/Japanese+Credit+Screening>
- iv. Iran - Private dataset, Contact Hassan Sabzevari, hn_sabzevari@yahoo.com
- v. Spain - Contact Manuel Artis, manuel.artis@ub.edu
- vi. UCSD - Originally accessed at <http://mll.ucsd.edu/>, however the original dataset is no longer publicly available for download. A copy of the dataset is

available from the author at kennedykenneth@gmail.com.

vii. PAKDD - Competition dataset accessed at <http://sede.neurotech.com.br/>

PAKDD2009/

viii. Thomas - The data came as a cd-rom in Thomas *et al.* (2002)

ix. Poland - Competition dataset accessed at [http://www.pietruszkiewicz.](http://www.pietruszkiewicz.com/index.php?main=dataset)

[com/index.php?main=dataset](http://www.pietruszkiewicz.com/index.php?main=dataset)

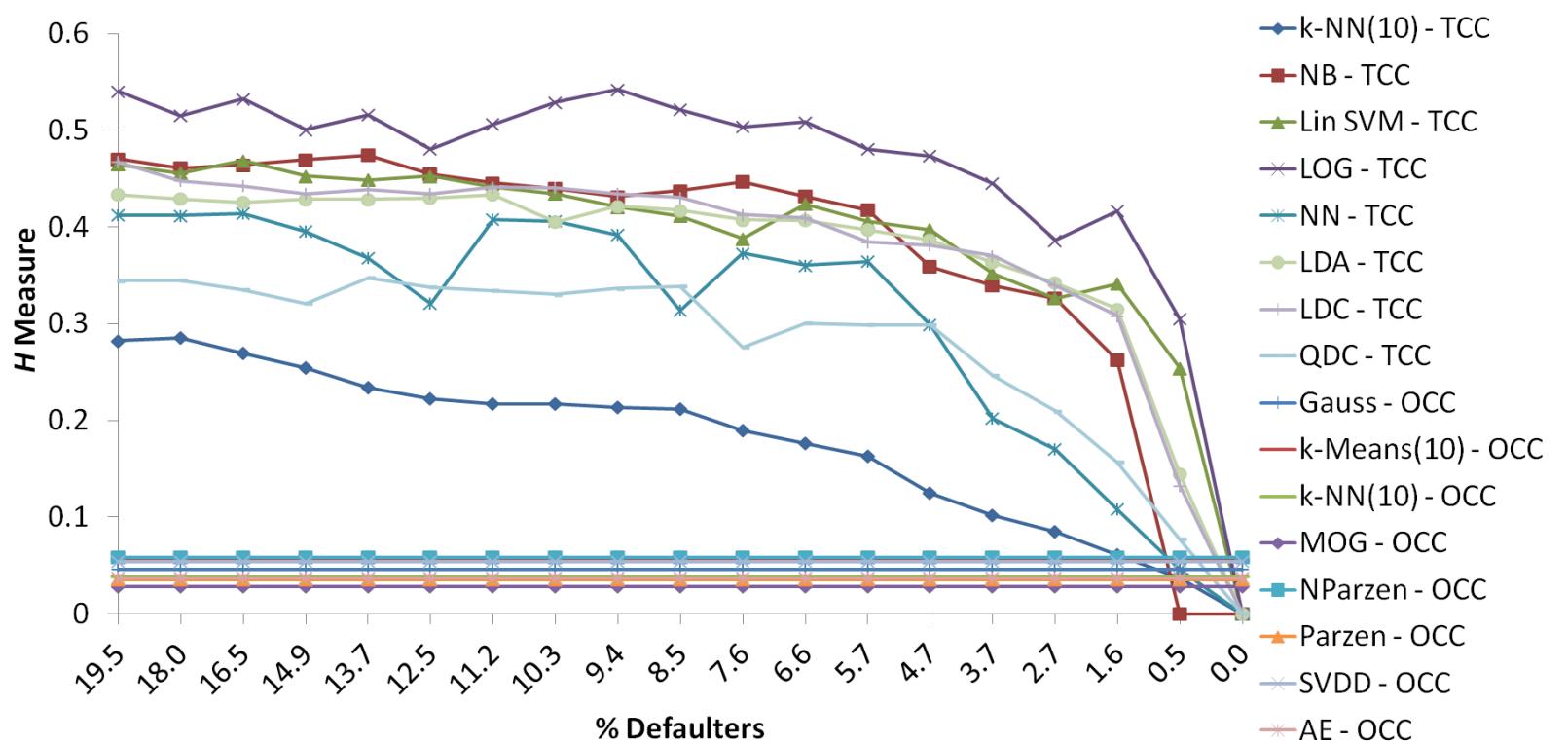


Figure C.1: Iran: Normal process and one-class classification process test set H measure performance.

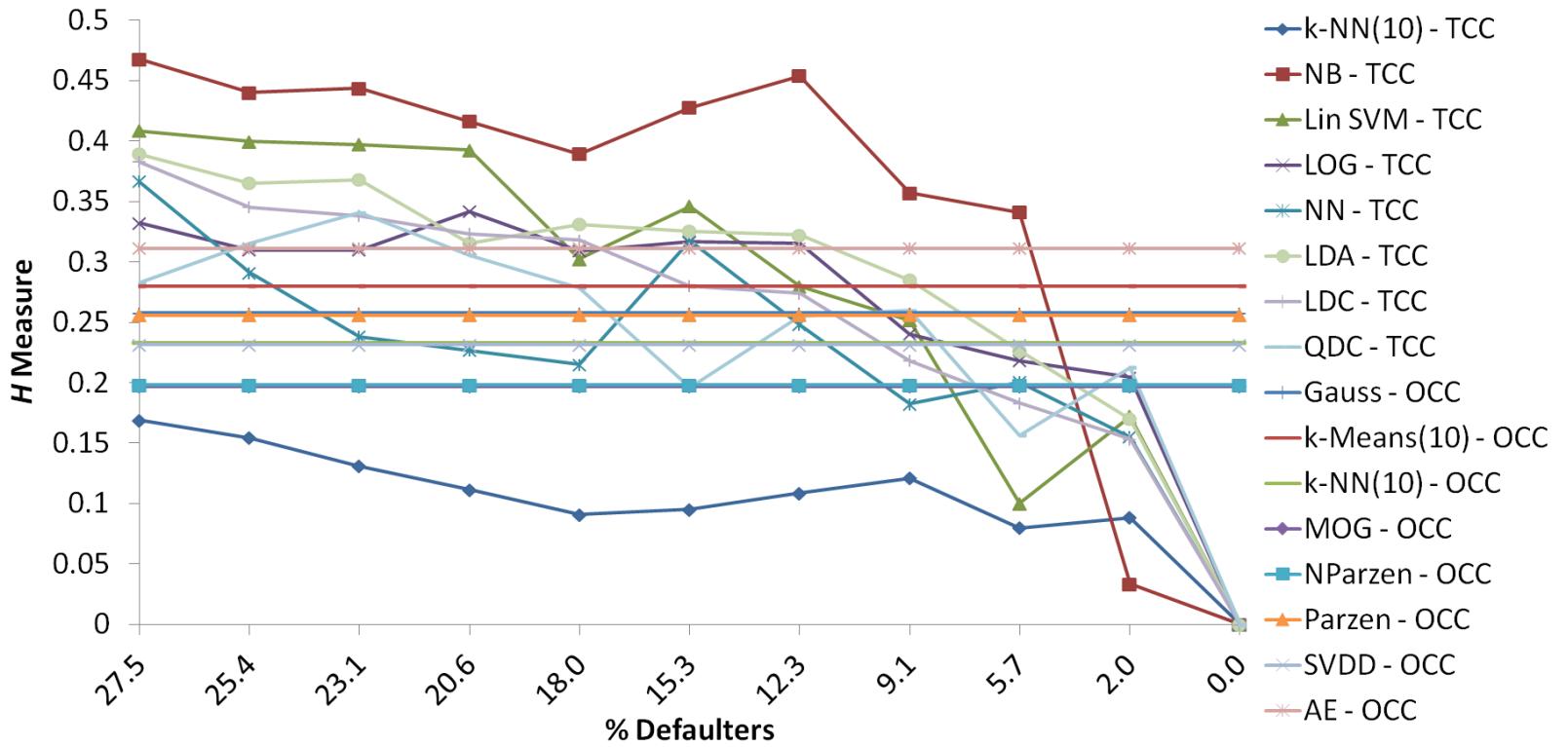


Figure C.2: Japan: Normal process and one-class classification process test set H measure performance.

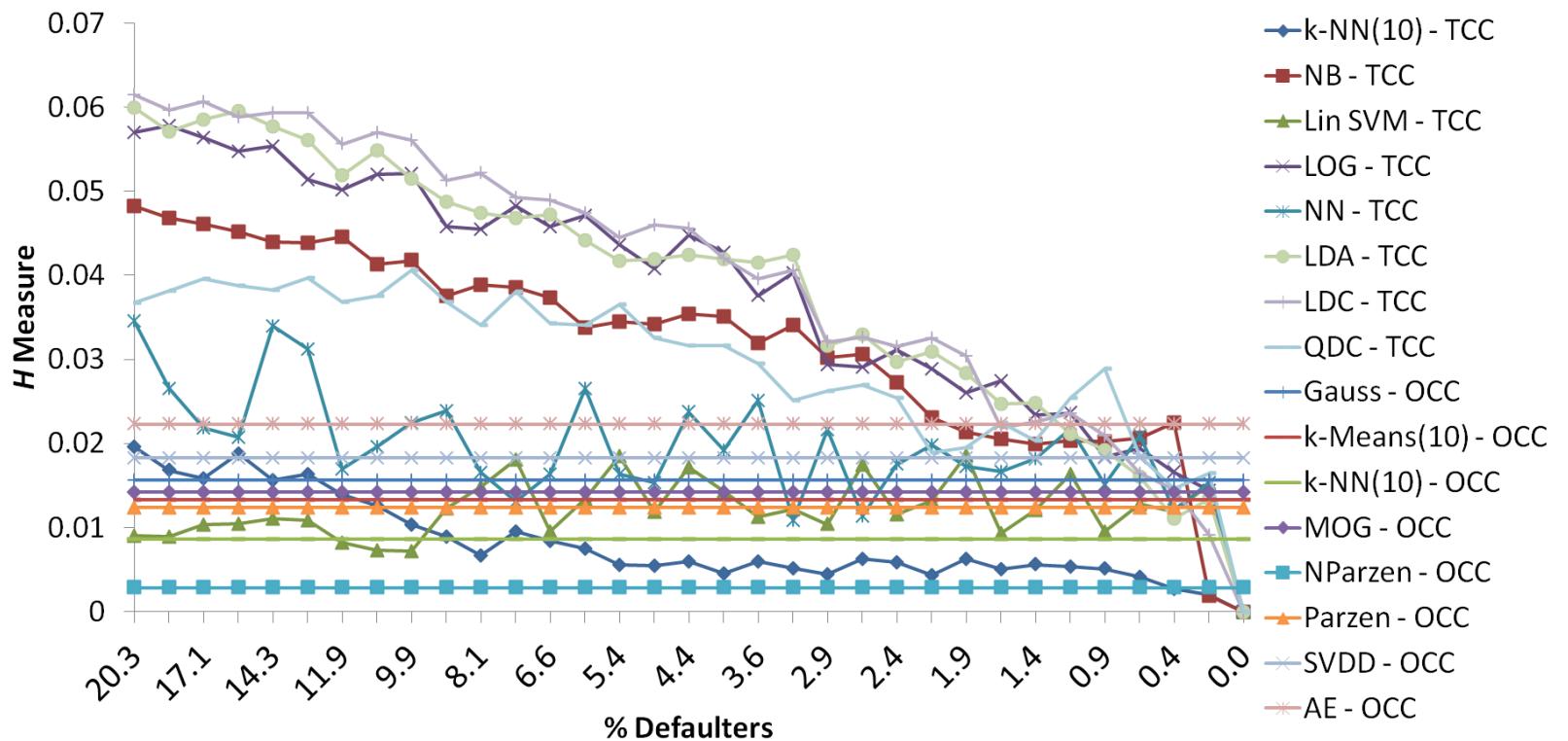


Figure C.3: PAKDD: Normal process and one-class classification process test set H measure performance.

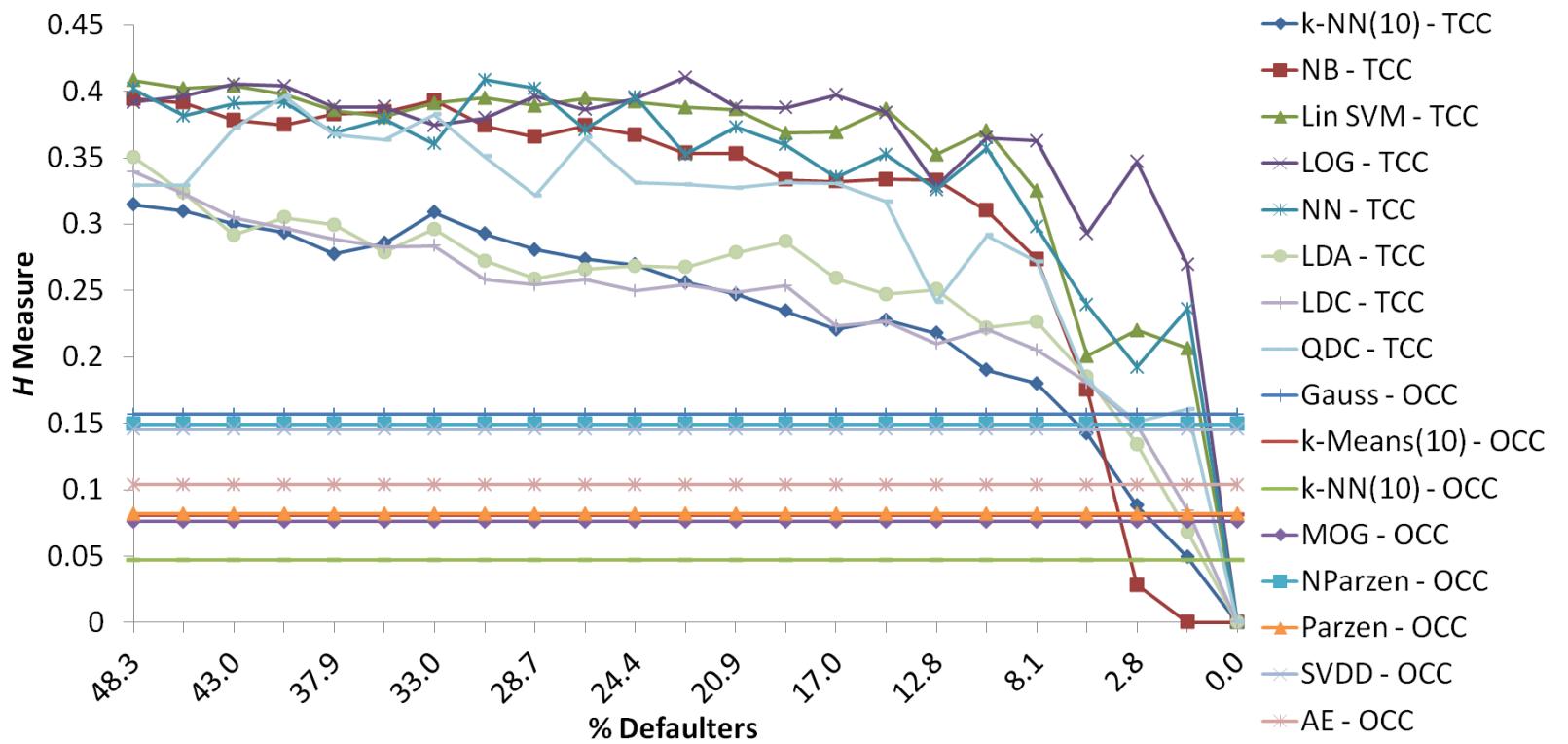


Figure C.4: Poland: Normal process and one-class classification process test set H measure performance.

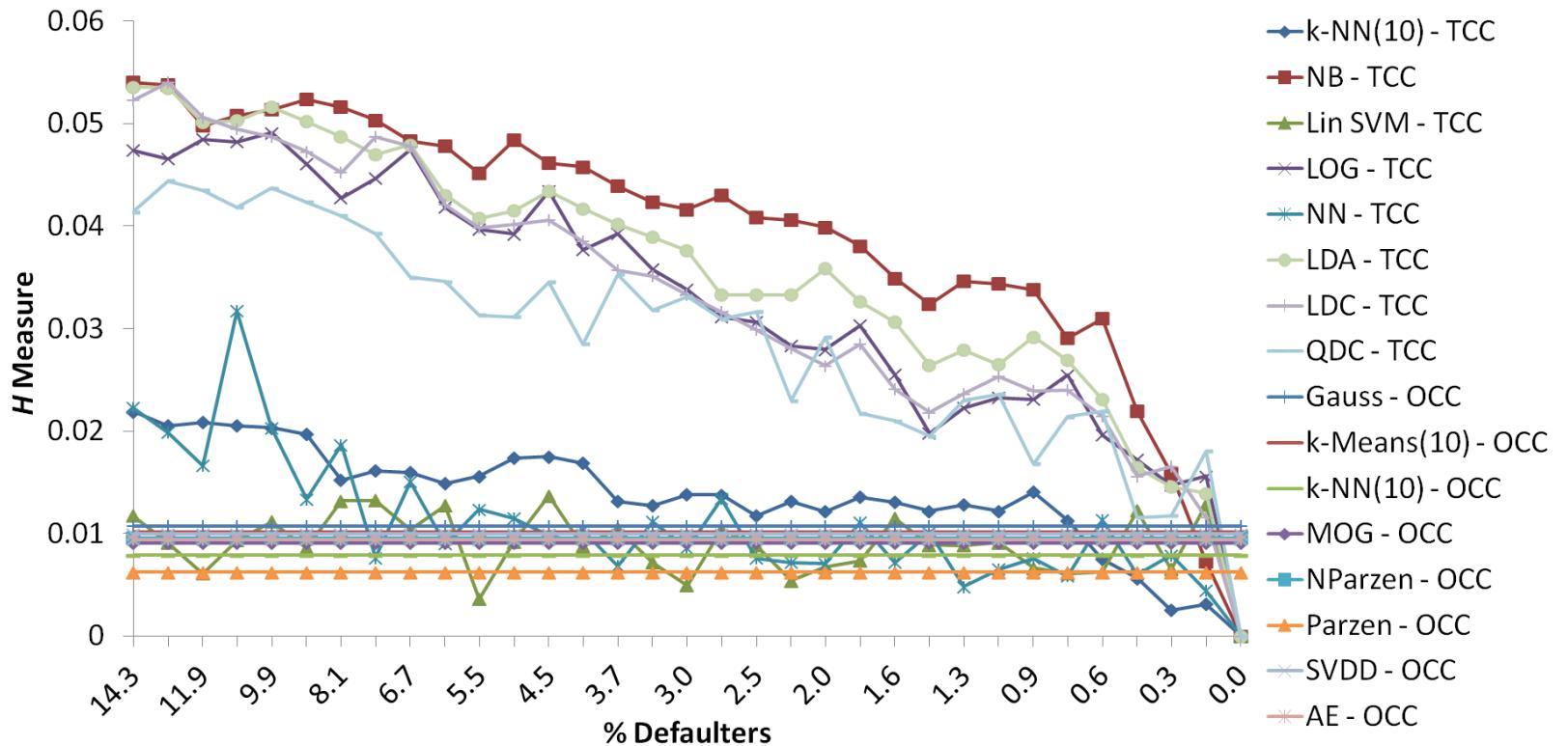


Figure C.5: Spain: Normal process and one-class classification process test set H measure performance.

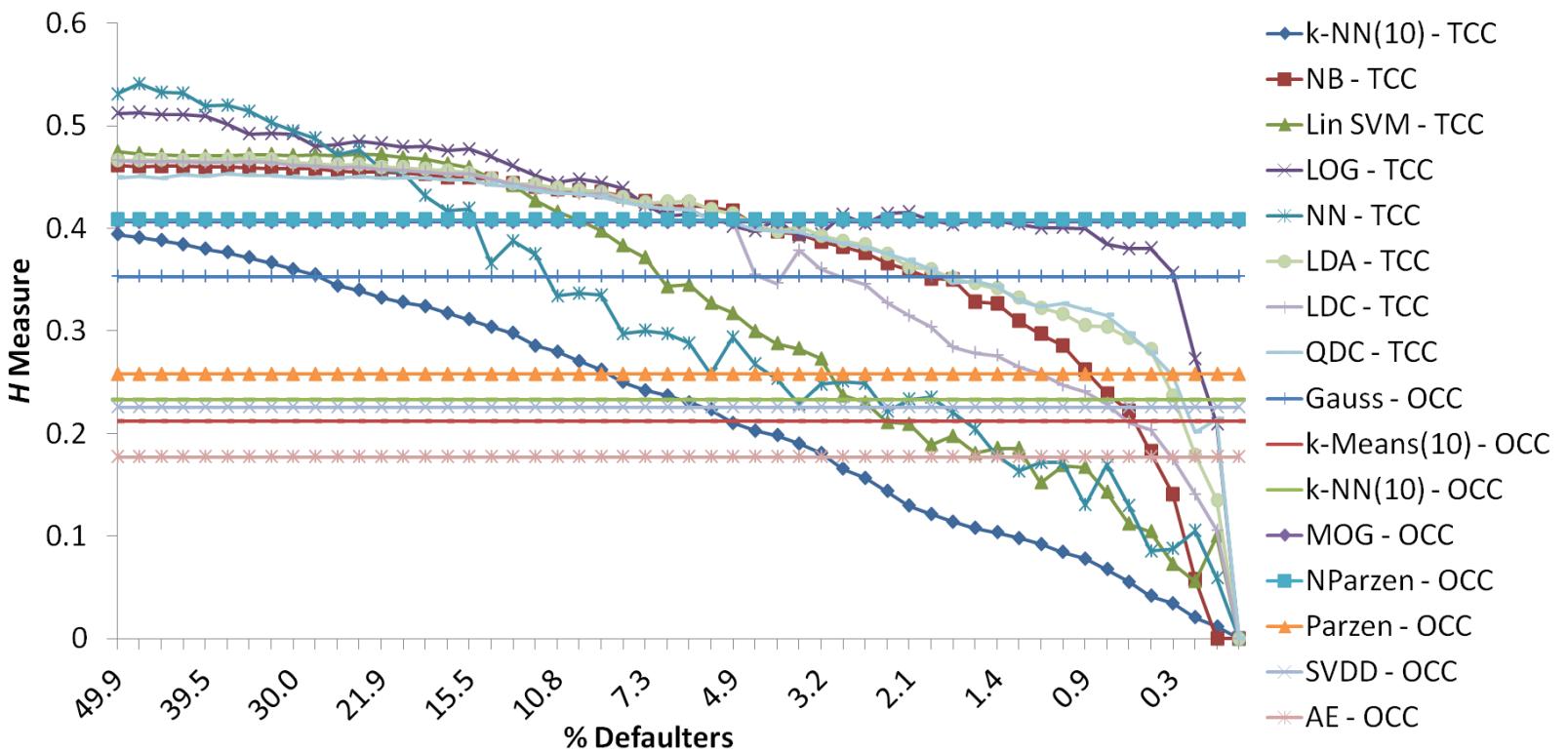


Figure C.6: UCSD: Normal process and one-class classification process test set H measure performance.

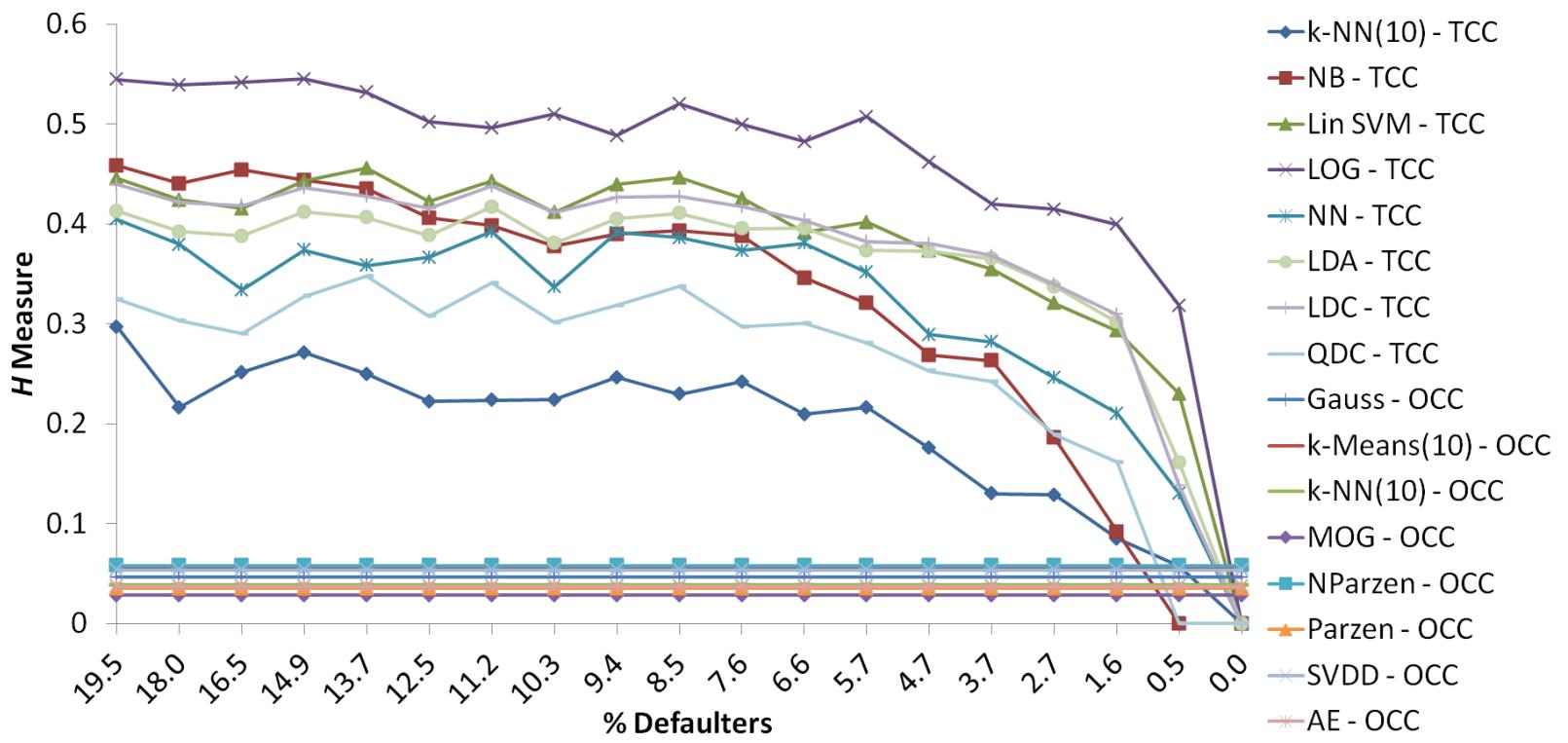


Figure C.7: Iran: Oversample process and one-class classification process test set H measure performance.

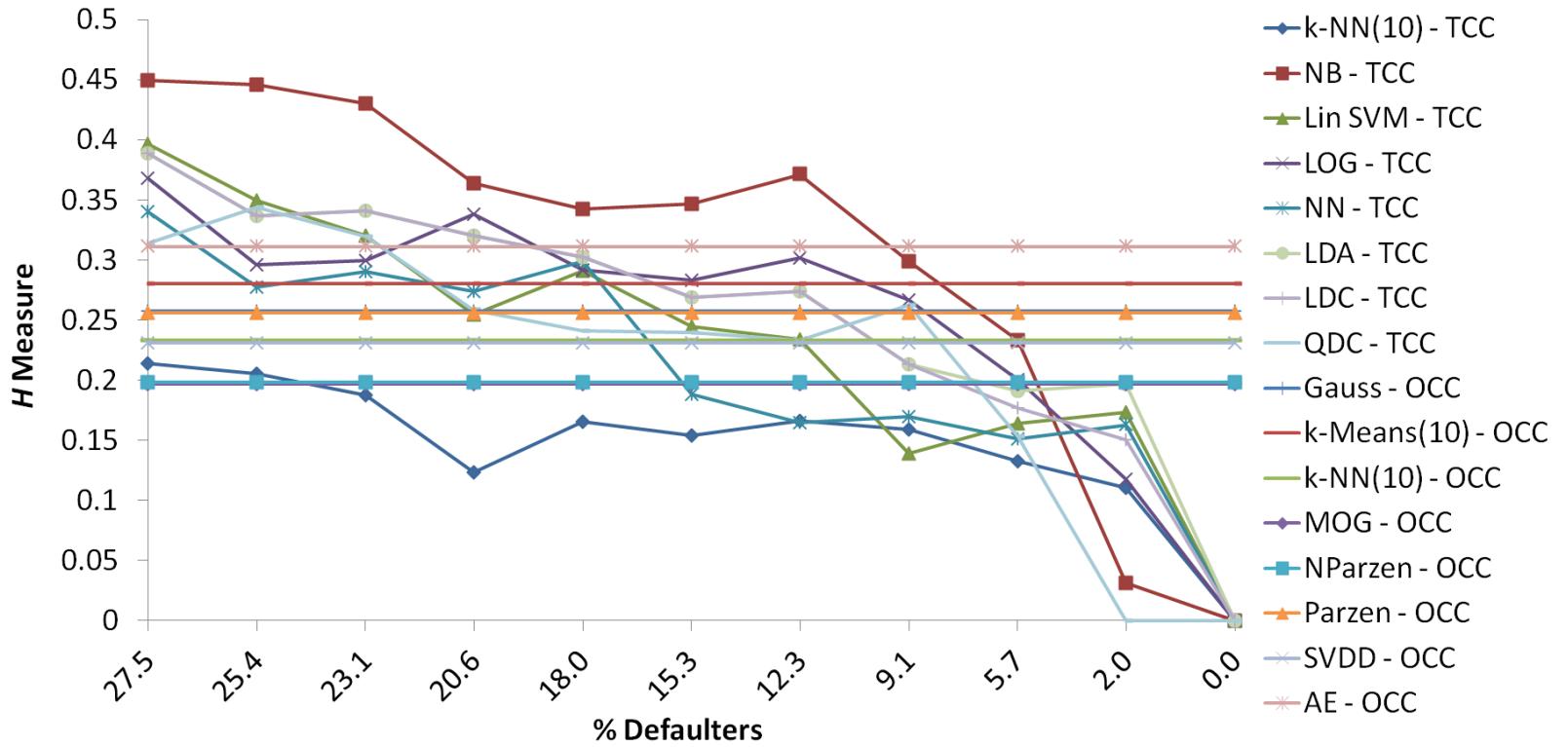


Figure C.8: Japan: Oversample process and one-class classification process test set H measure performance.

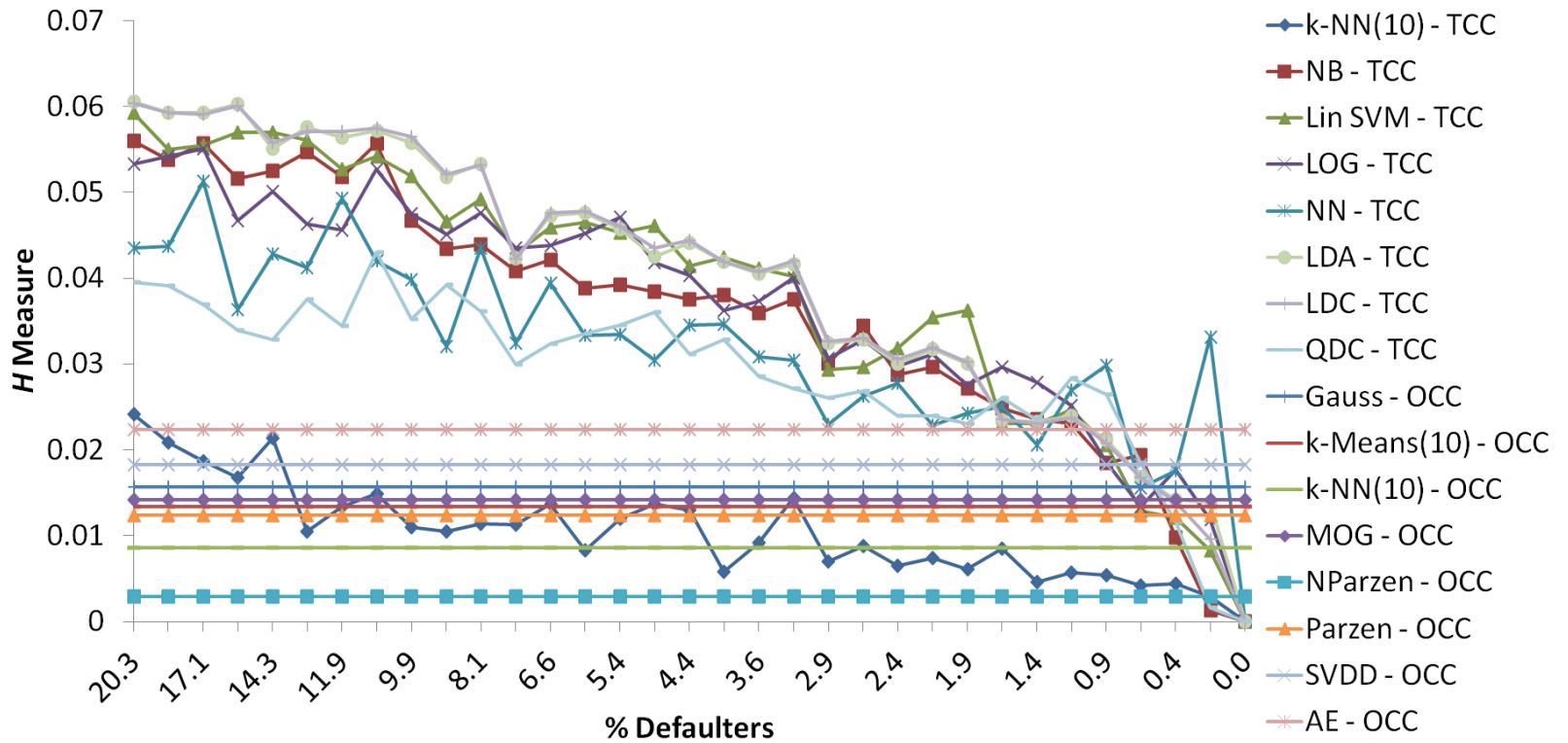


Figure C.9: PAKDD: Oversample process and one-class classification process test set H measure performance.

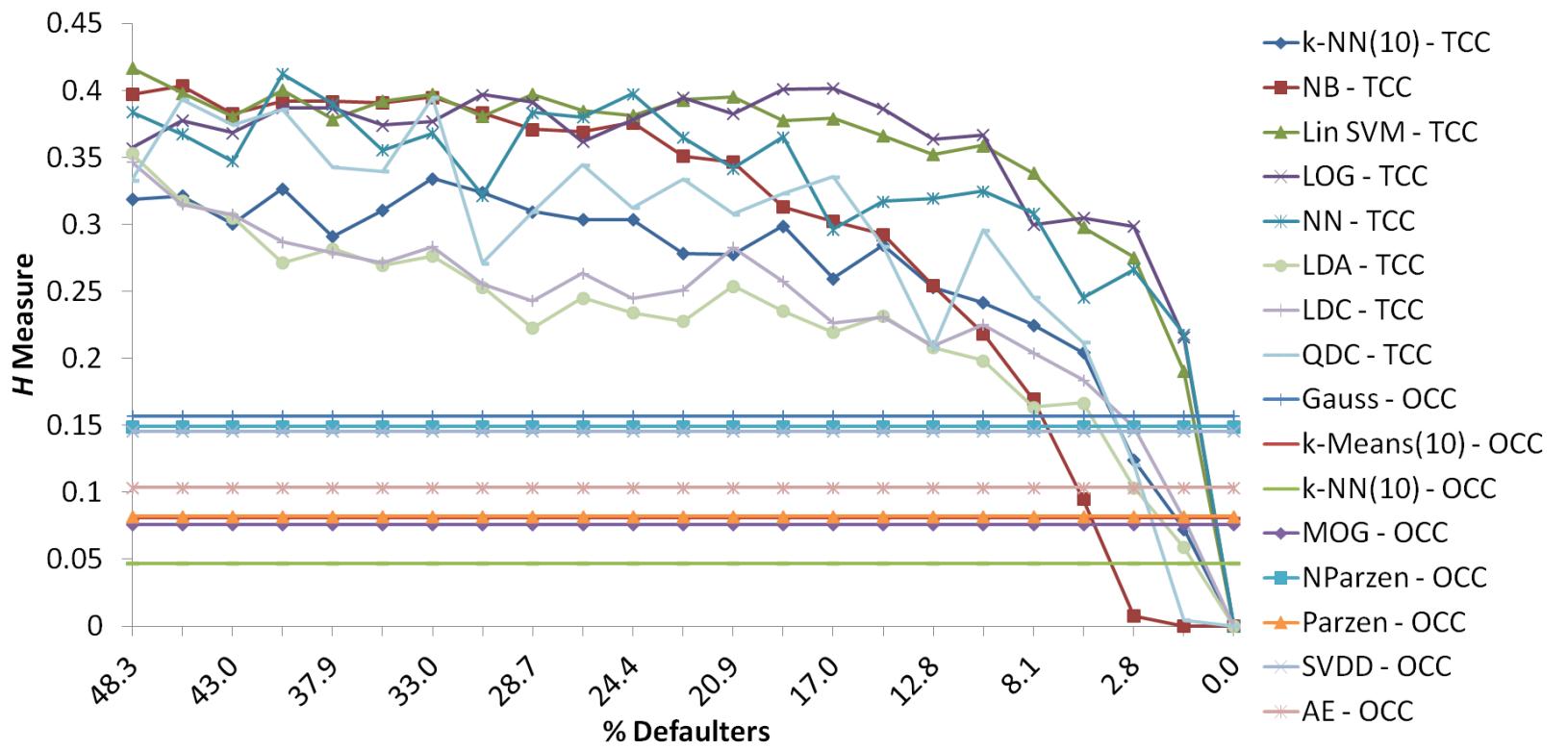


Figure C.10: Poland: Oversample process and one-class classification process test set H measure performance.

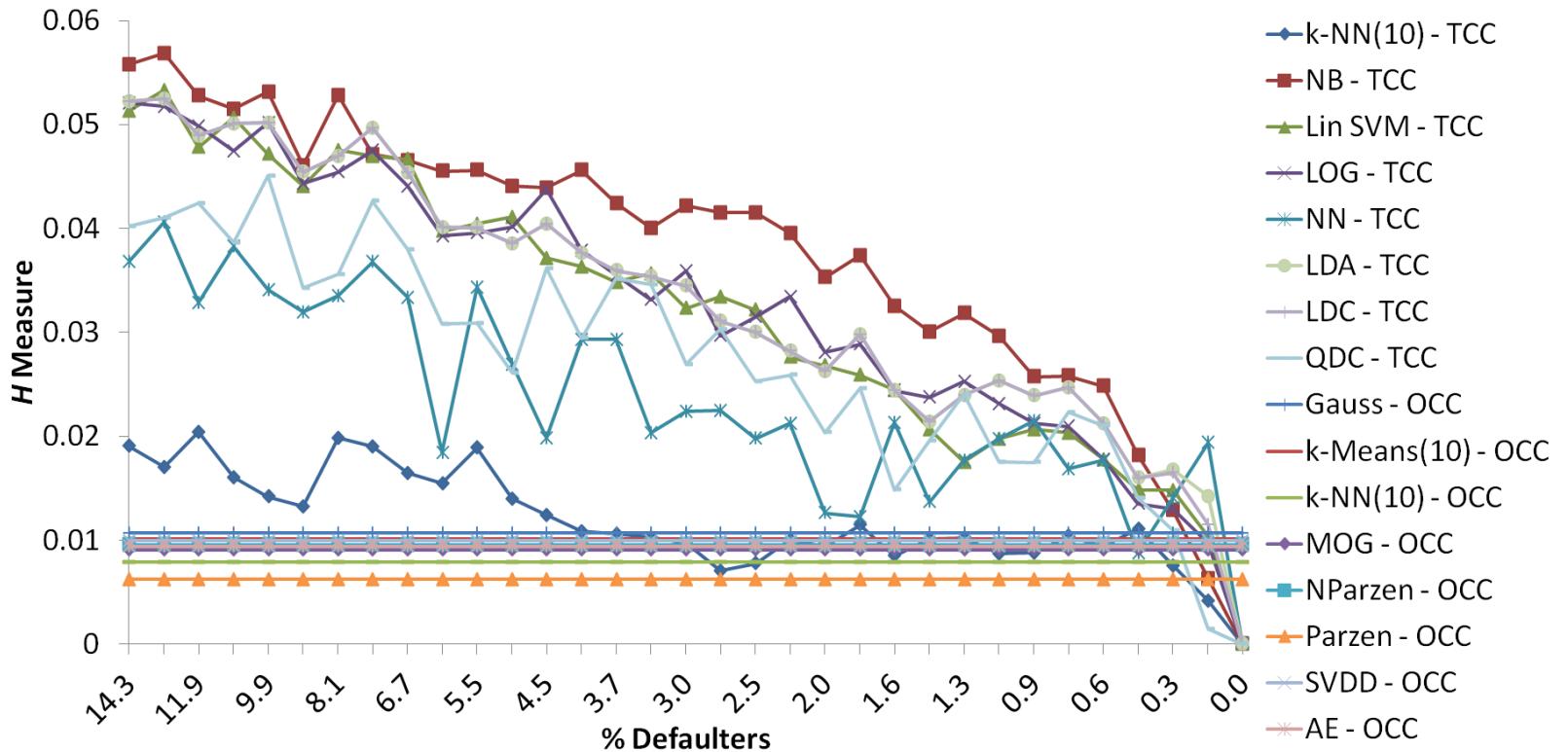


Figure C.11: Spain: Oversample process and one-class classification process test set H measure performance.

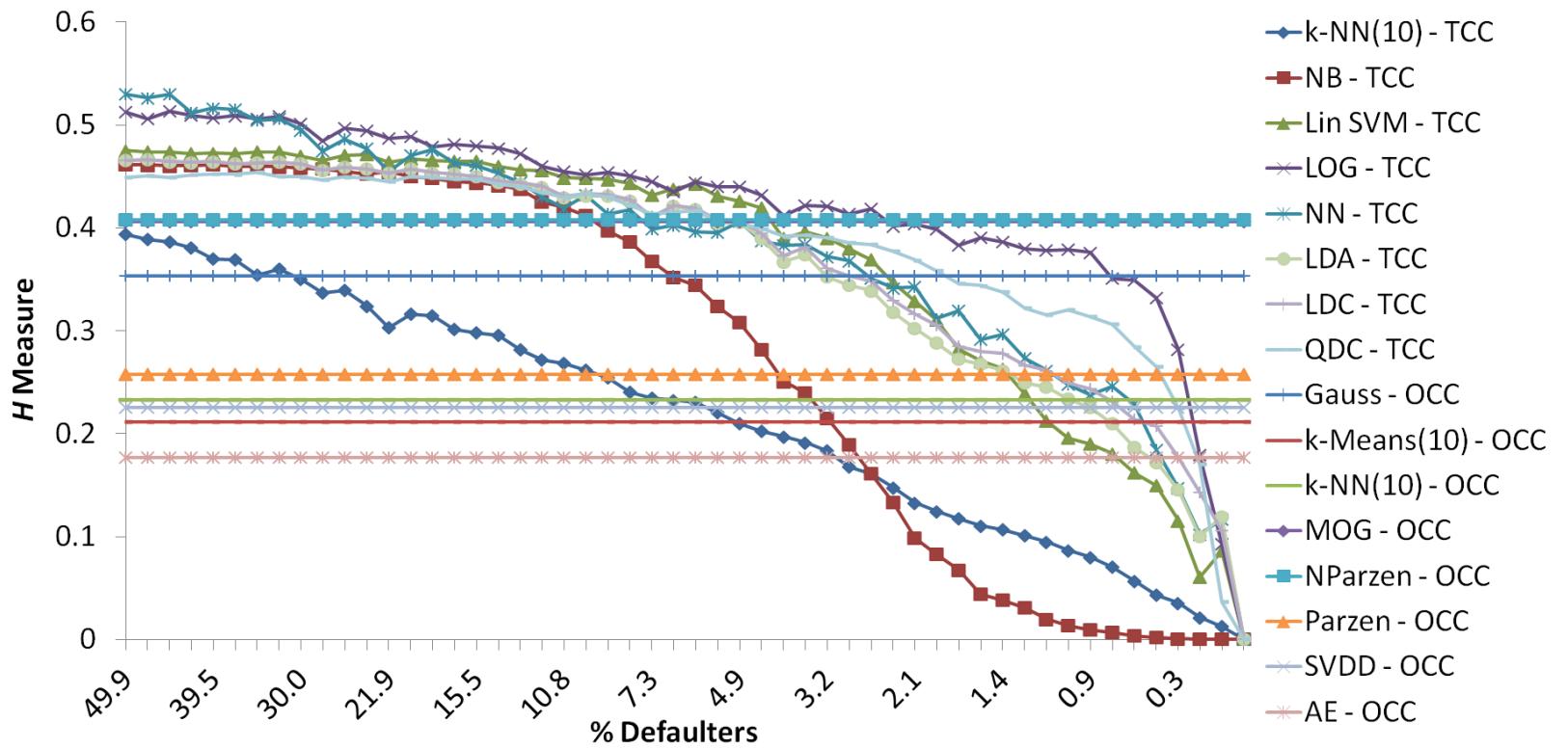


Figure C.12: UCSD: Oversample process and one-class classification process test set H measure performance.

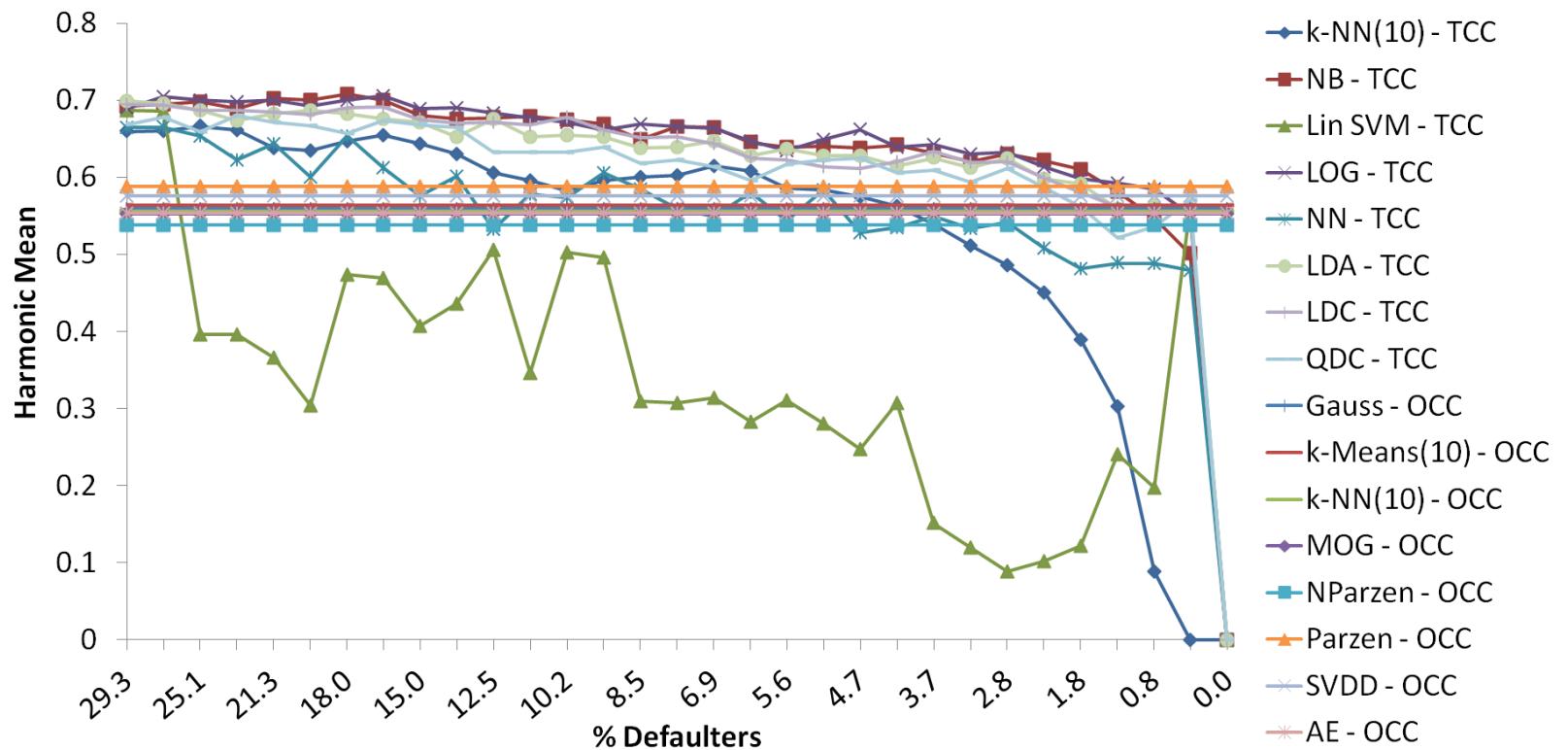


Figure C.13: German: Normal process and one-class classification process test set harmonic mean performance.

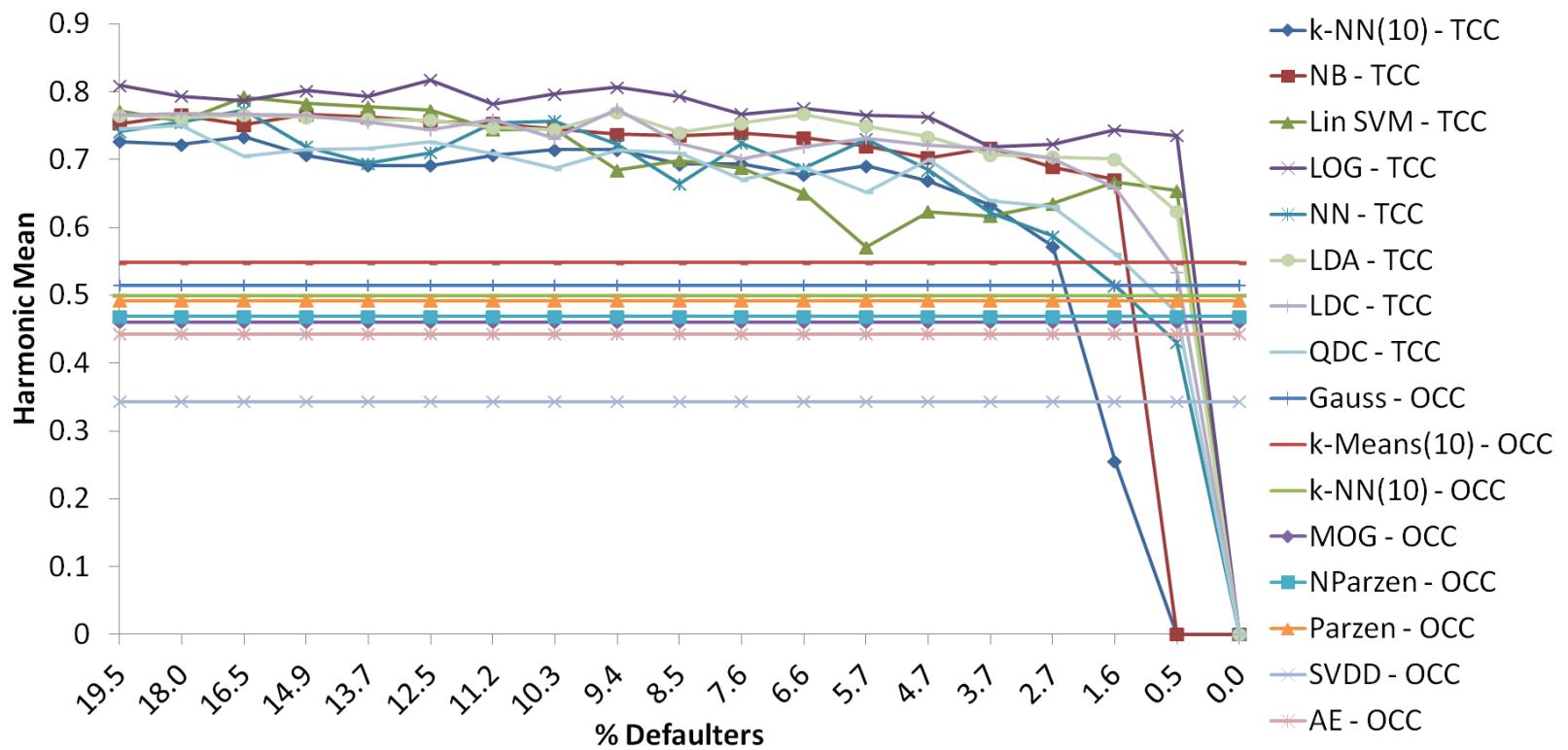


Figure C.14: Iran: Normal process and one-class classification process test set harmonic mean performance.

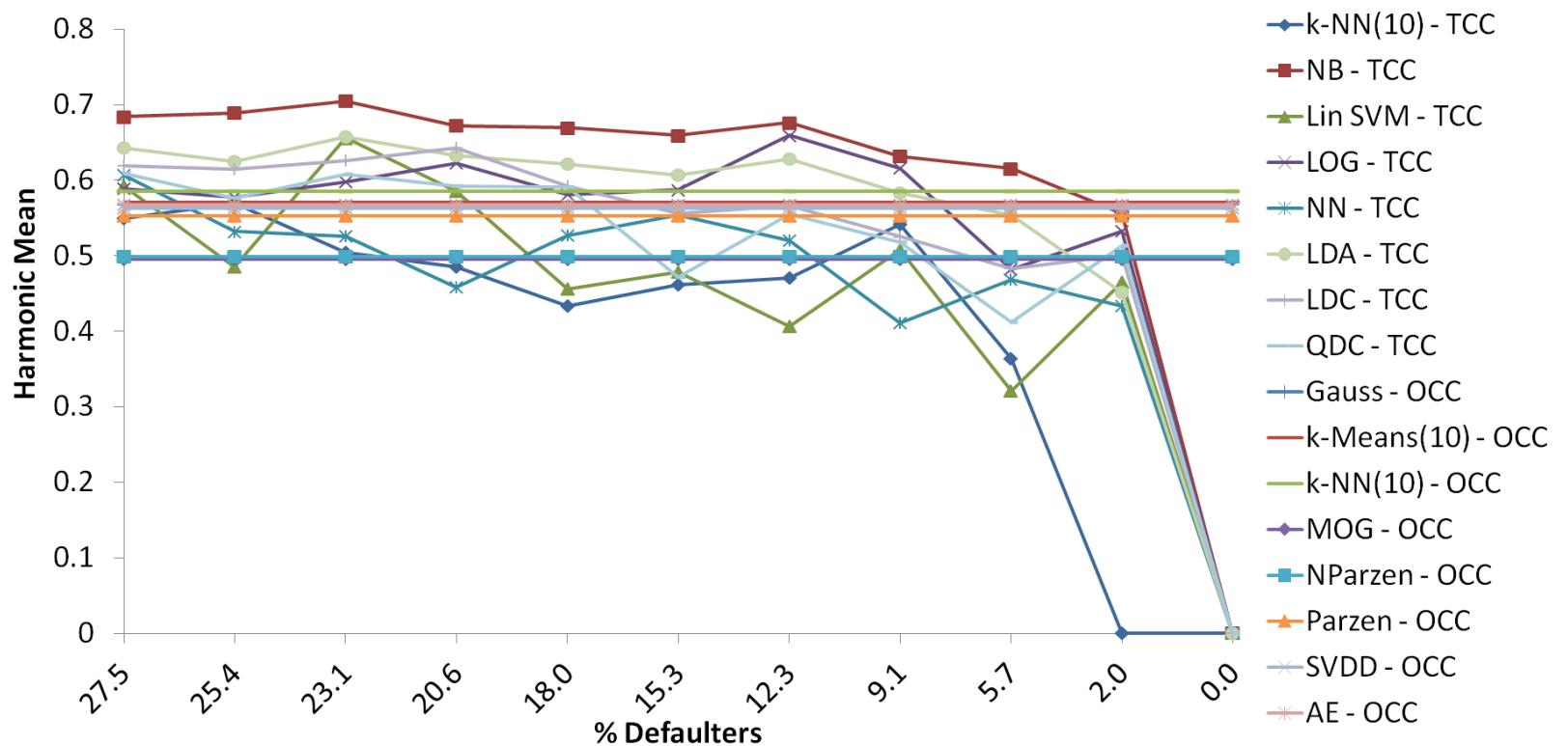


Figure C.15: Japan: Normal process and one-class classification process test set harmonic mean performance.

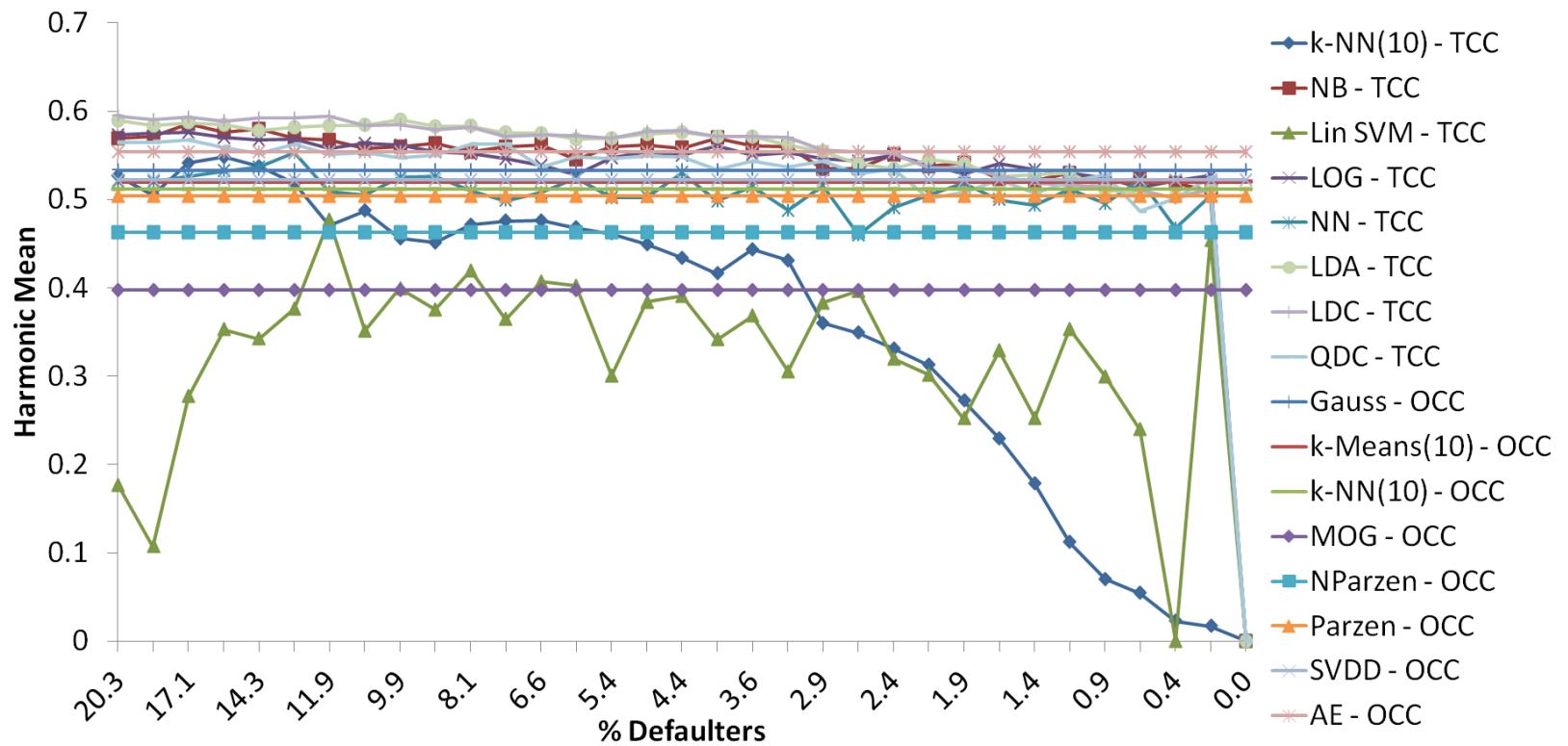


Figure C.16: PAKDD: Normal process and one-class classification process test set harmonic mean performance.

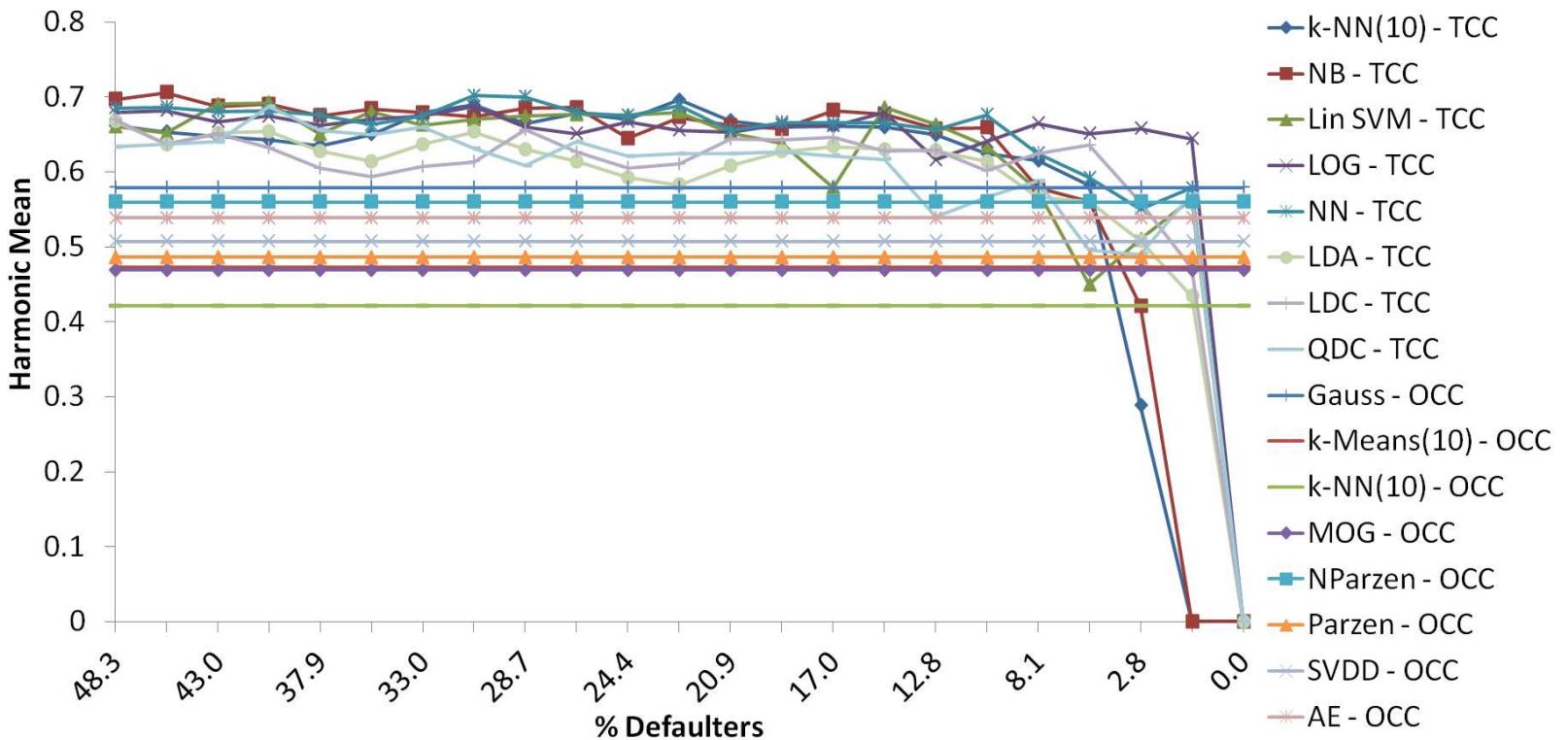


Figure C.17: Poland: Normal process and one-class classification process test set harmonic mean performance.

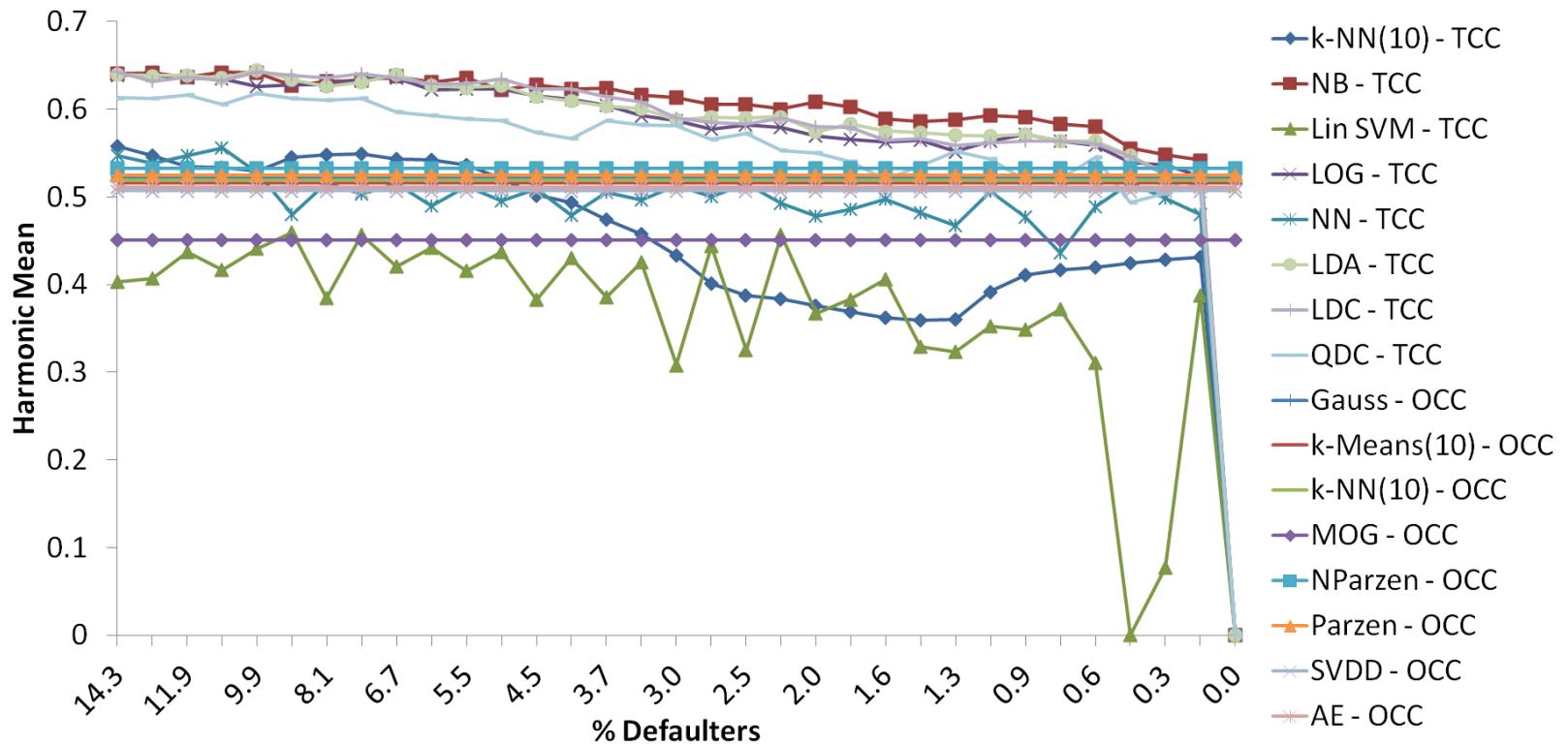


Figure C.18: Spain: Normal process and one-class classification process test set harmonic mean performance.

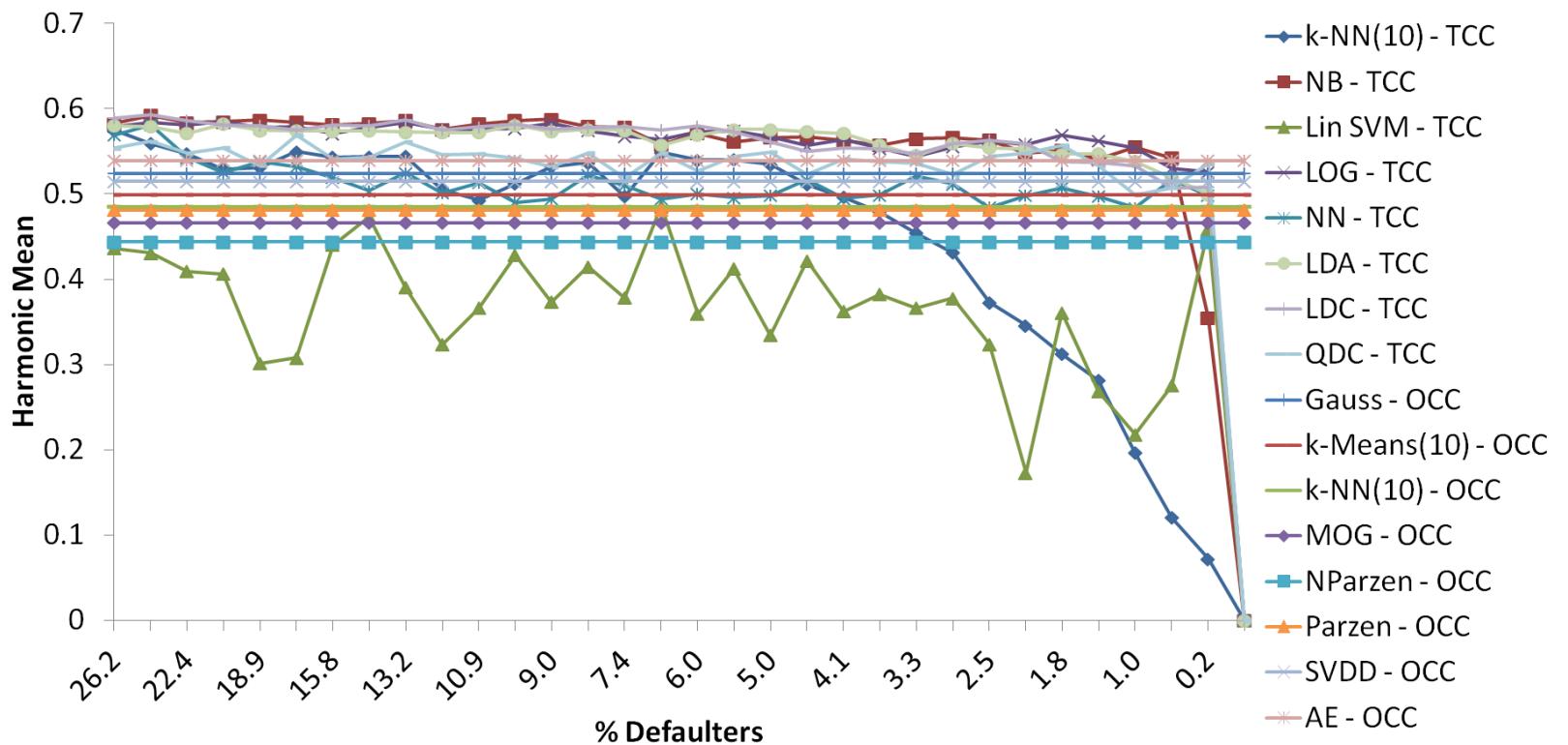


Figure C.19: Thomas: Normal process and one-class classification process test set harmonic mean performance.

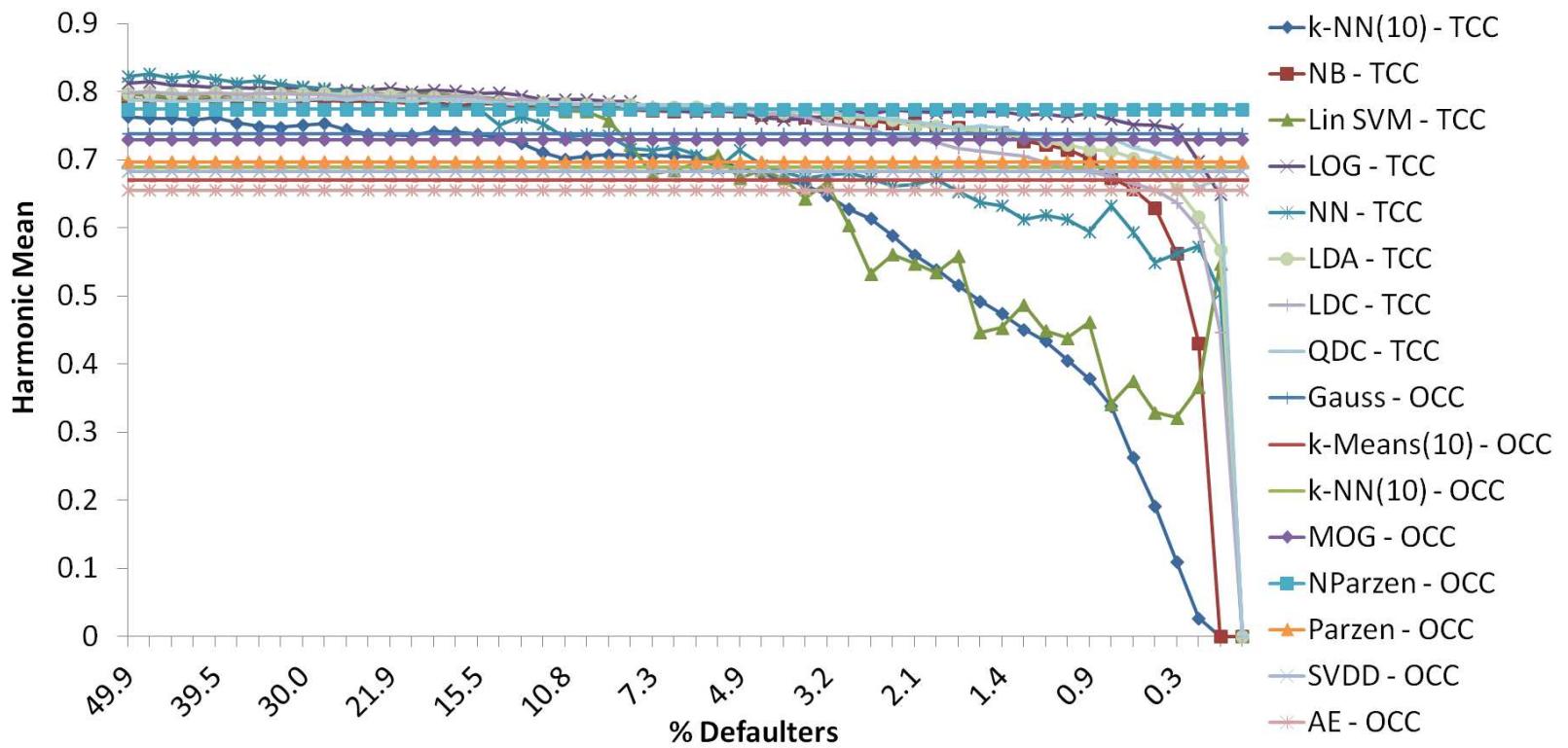


Figure C.20: UCSD: Normal process and one-class classification process test set harmonic mean performance.

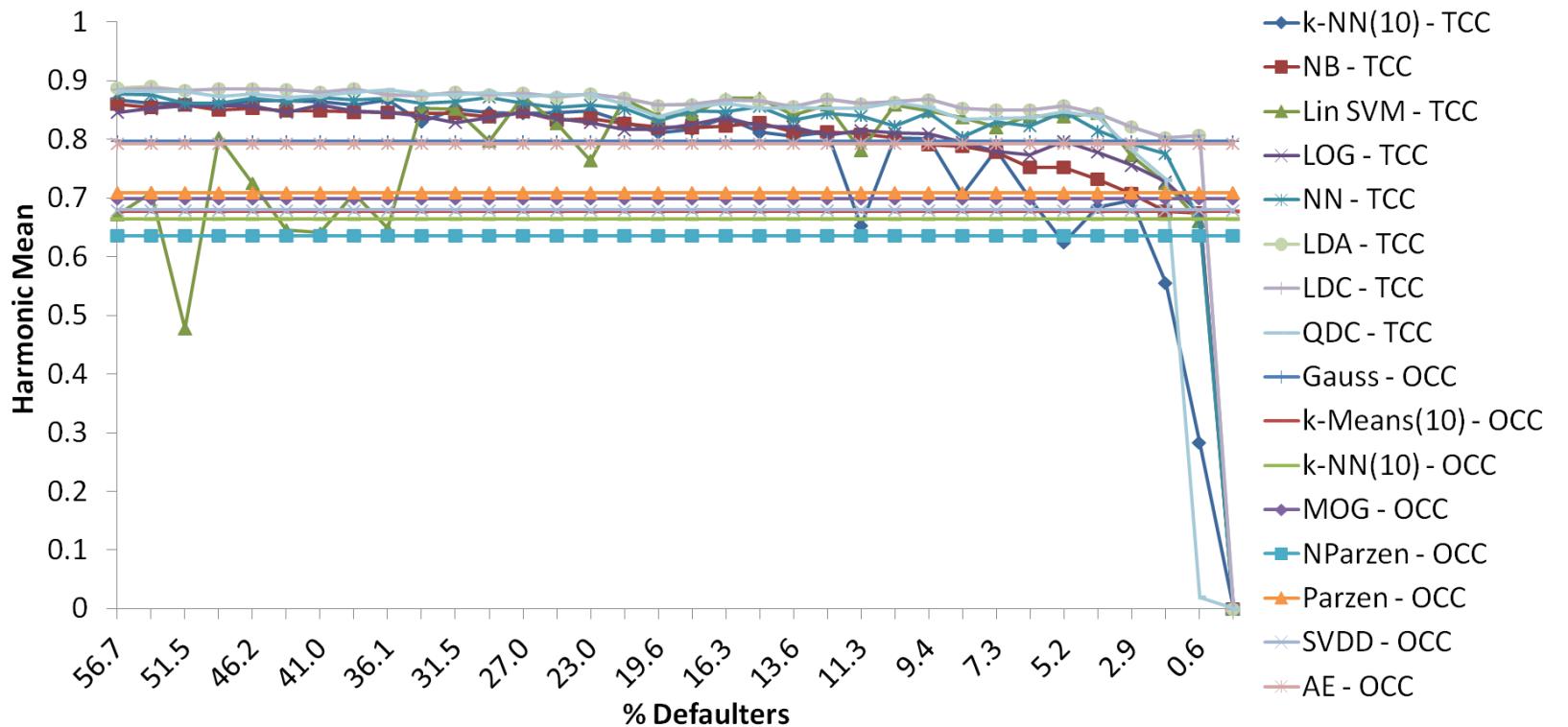


Figure C.21: Australia: Oversample process and one-class classification process test set harmonic mean performance.

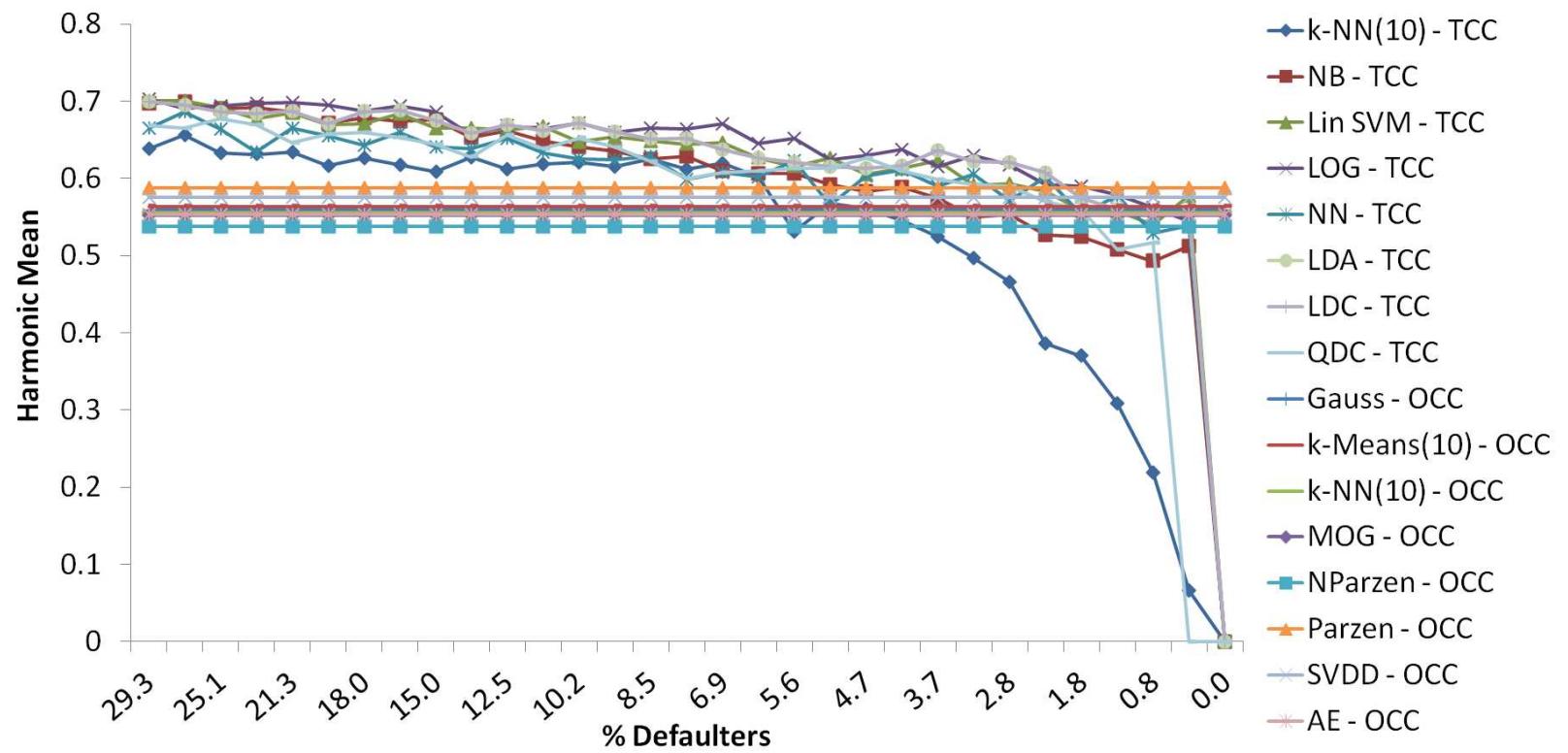


Figure C.22: German: Oversample process and one-class classification process test set harmonic mean performance.

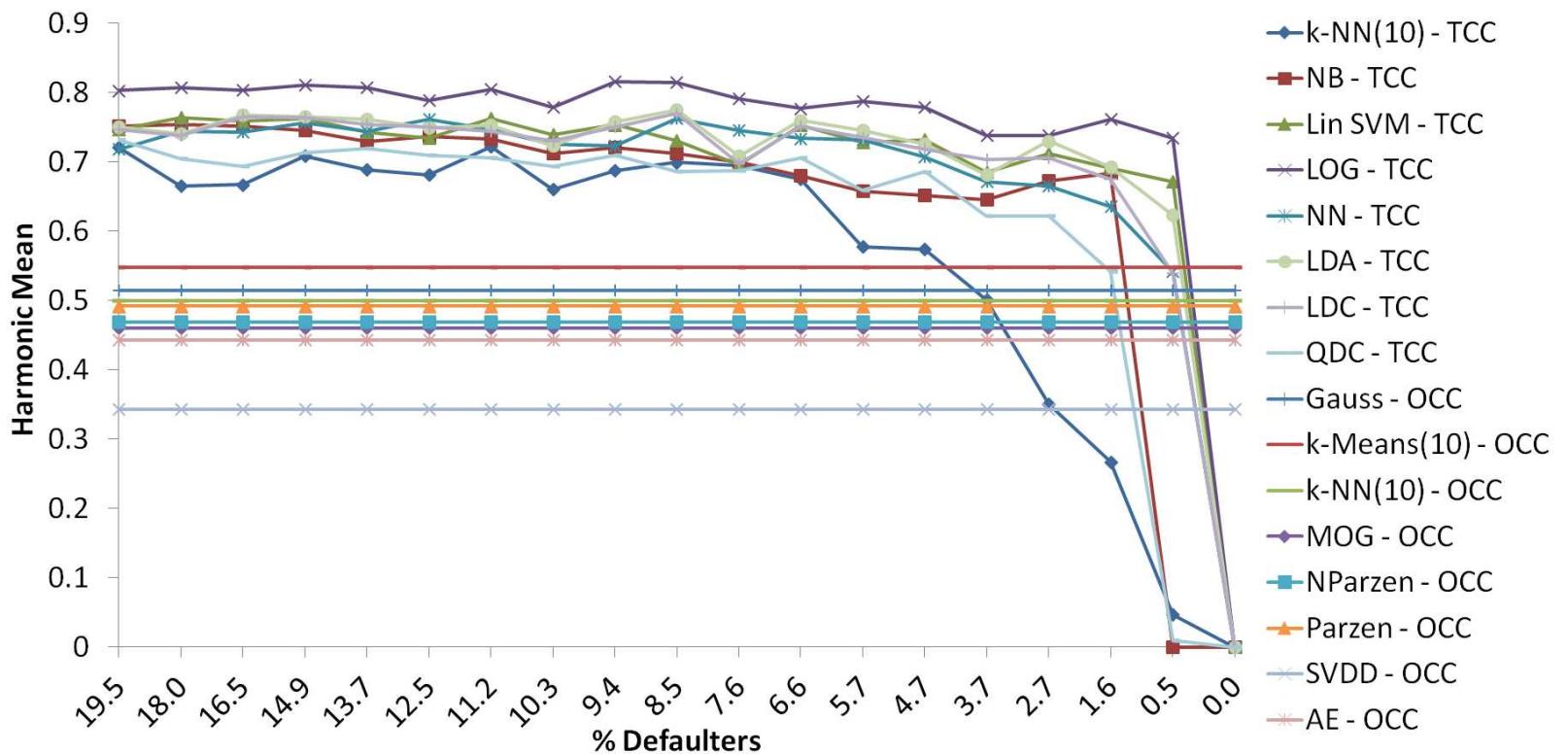


Figure C.23: Iran: Oversample process and one-class classification process test set harmonic mean performance.

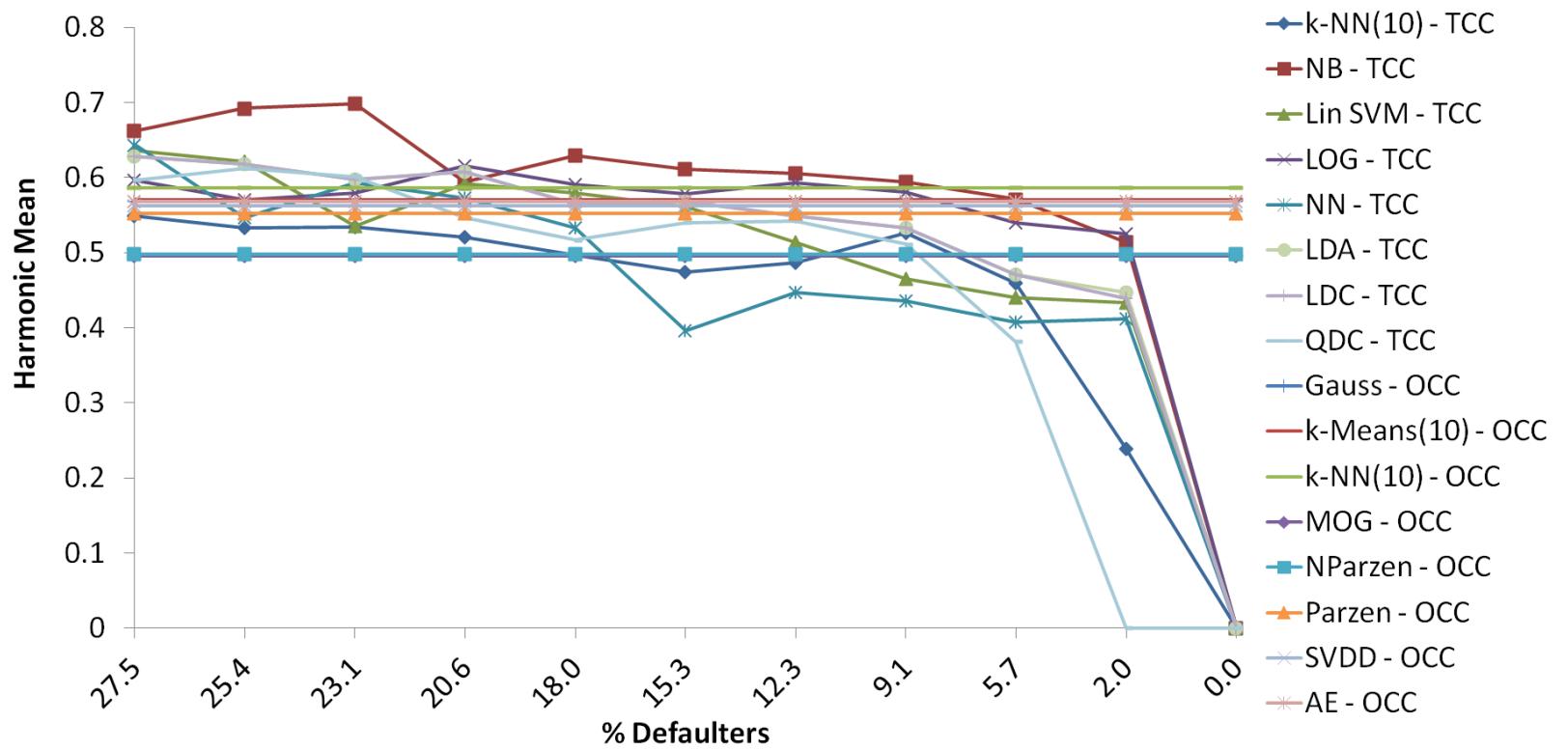


Figure C.24: Japan: Oversample process and one-class classification process test set harmonic mean performance.

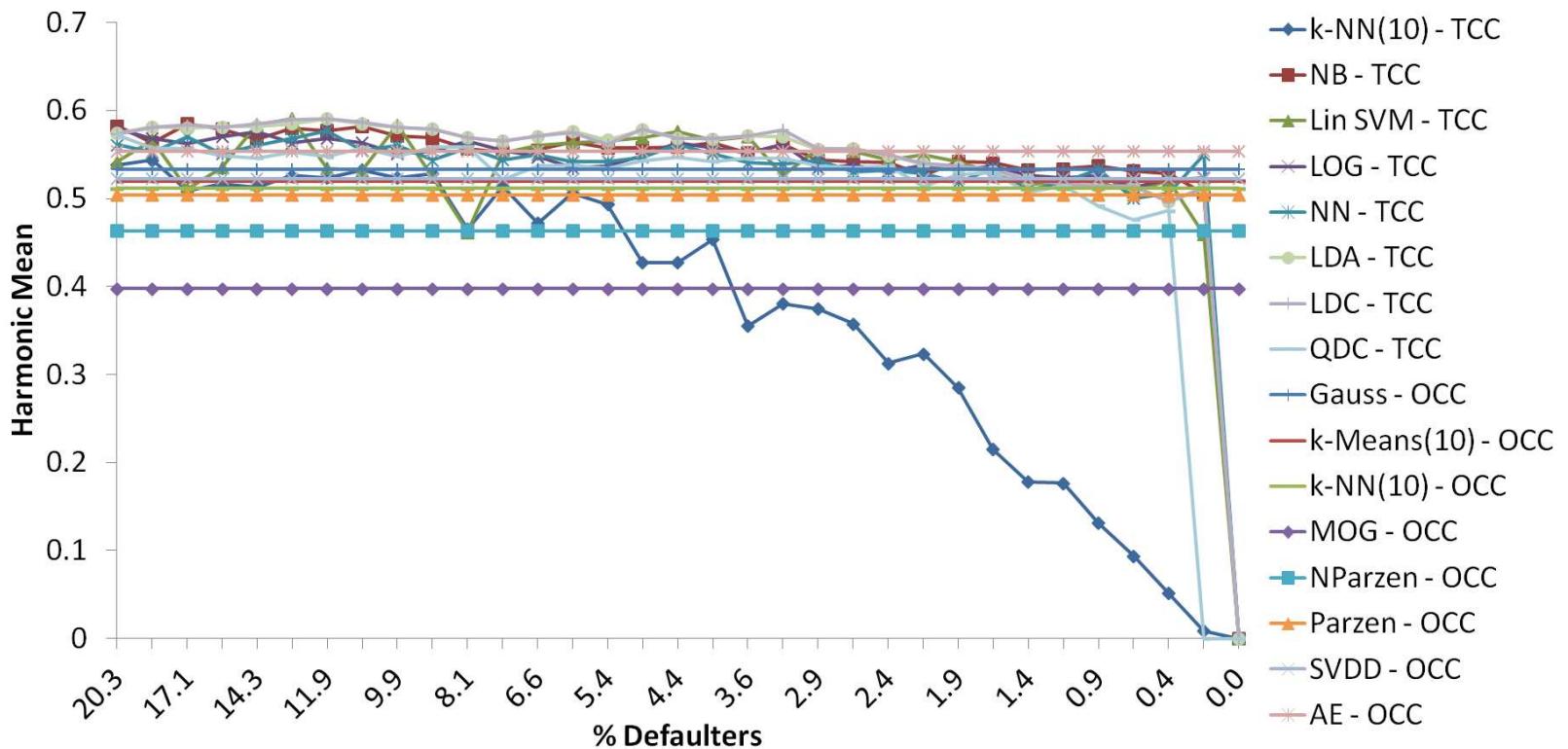


Figure C.25: PAKDD: Oversample process and one-class classification process test set harmonic mean performance.

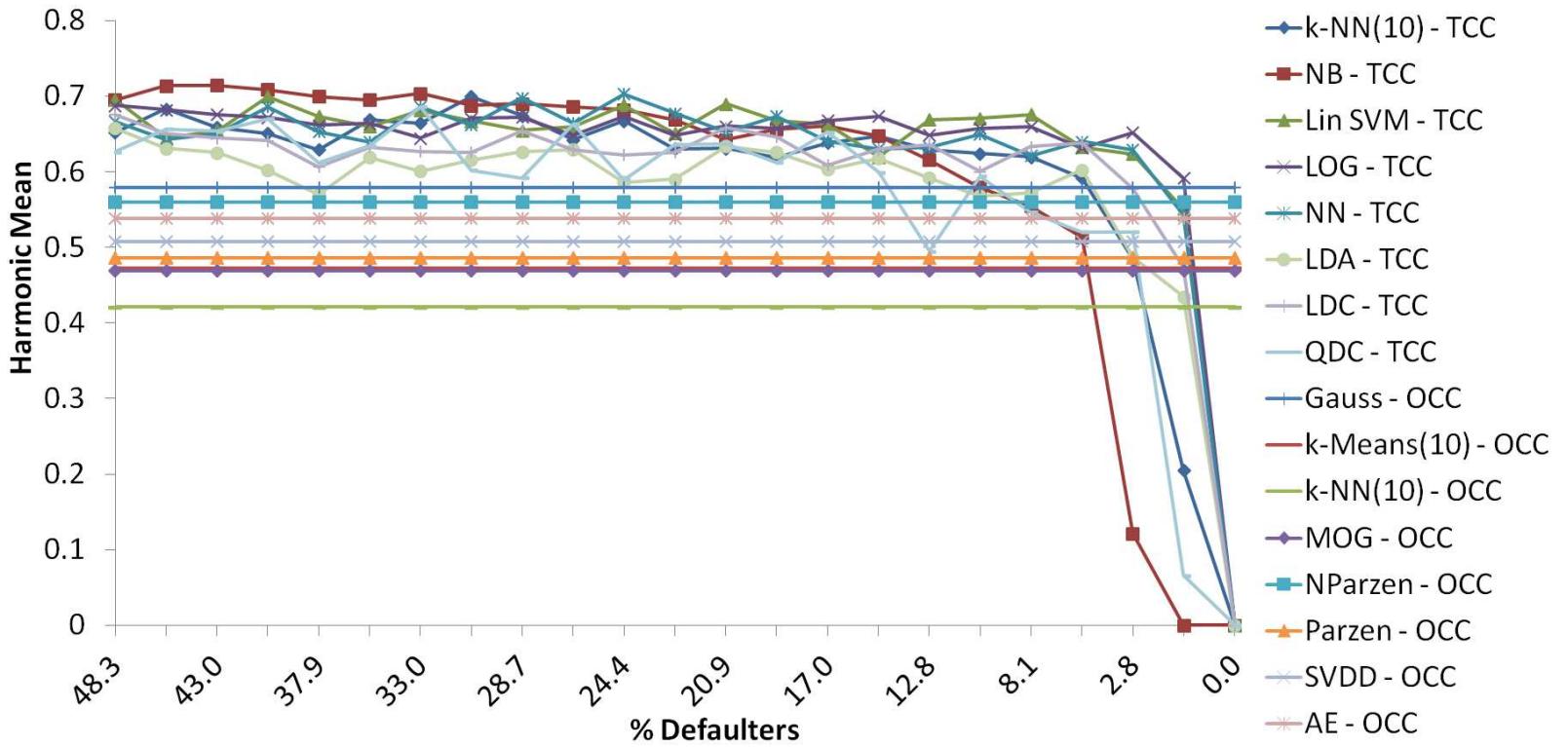


Figure C.26: Poland: Oversample process and one-class classification process test set harmonic mean performance.

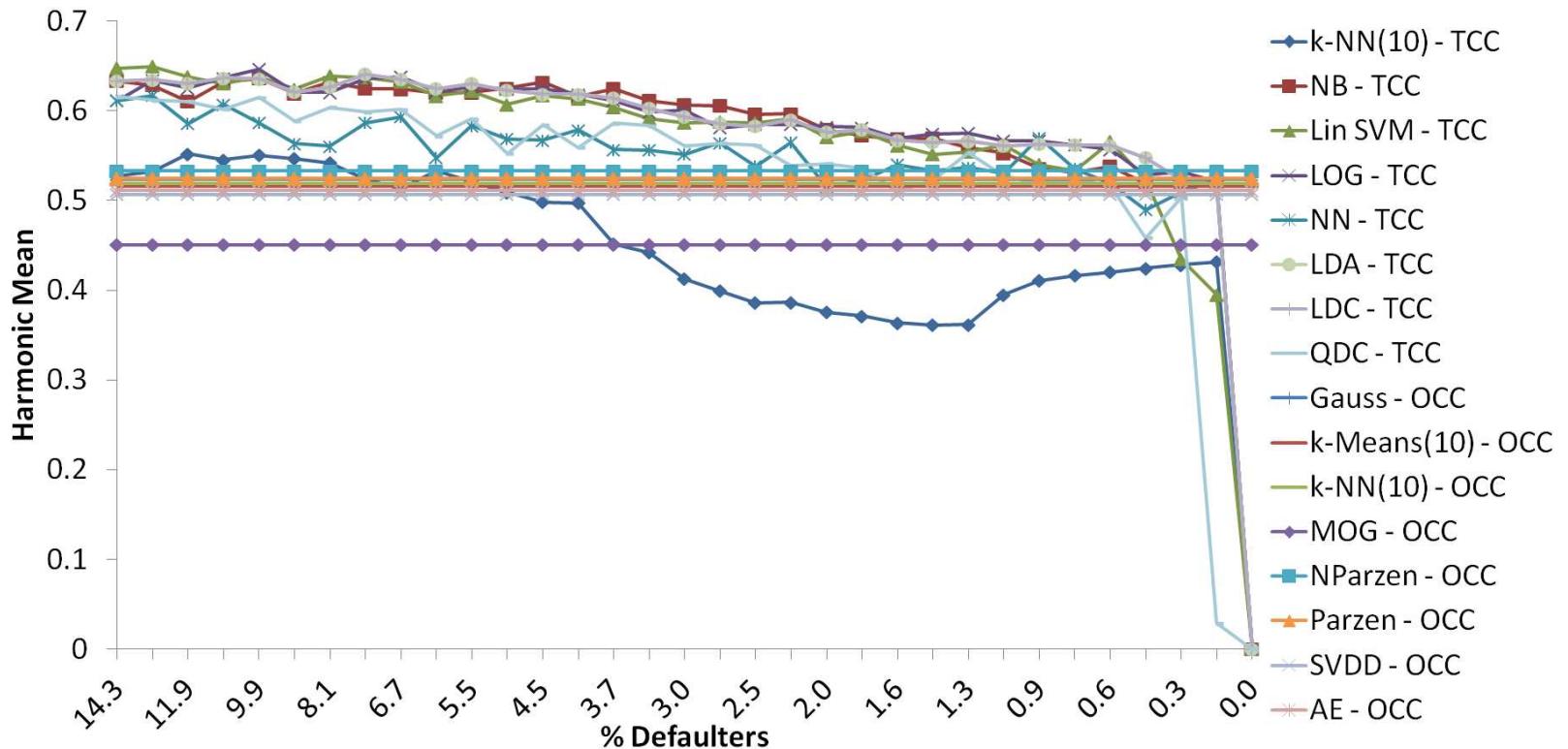


Figure C.27: Spain: Oversample process and one-class classification process test set harmonic mean performance.

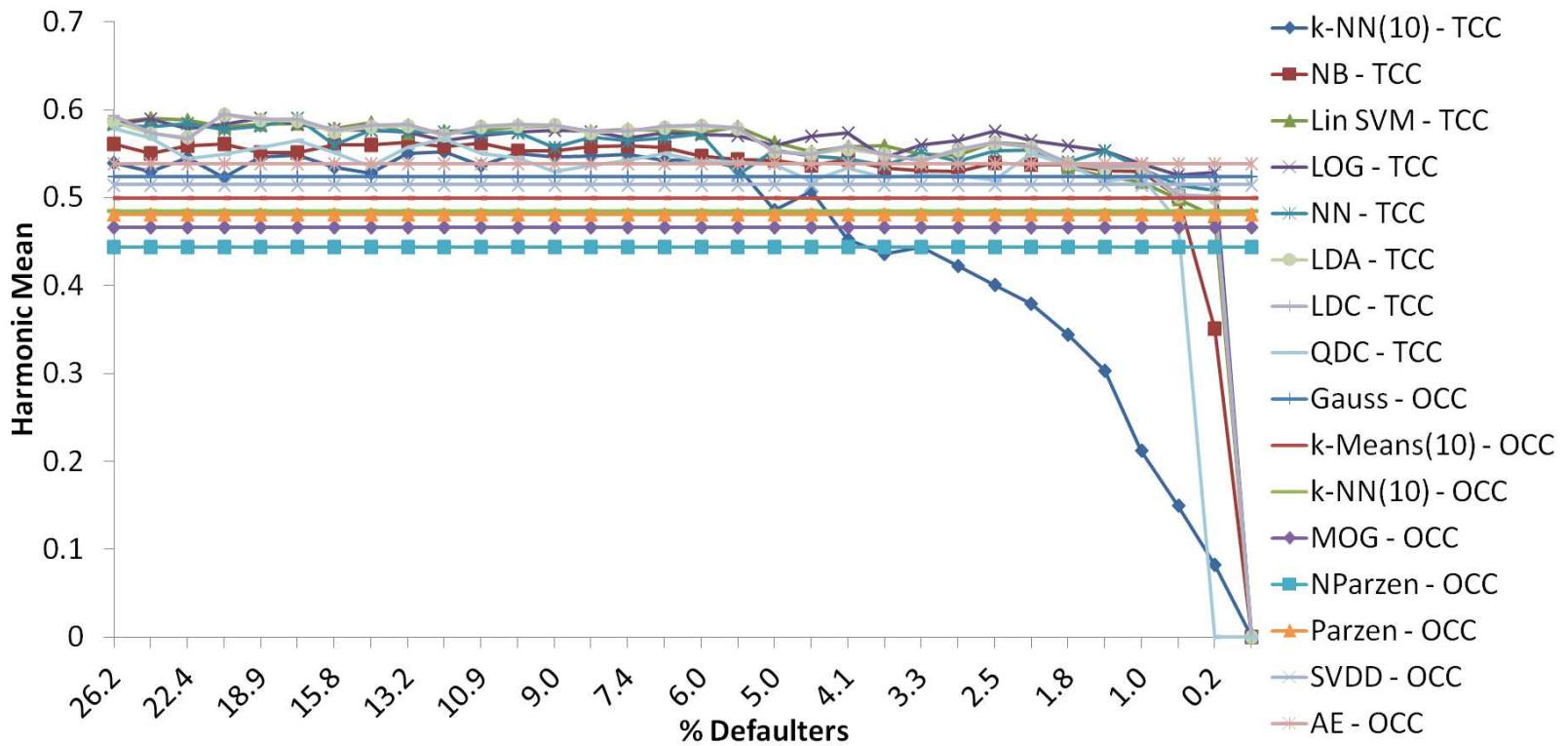


Figure C.28: Thomas: Oversample process and one-class classification process test set harmonic mean performance.

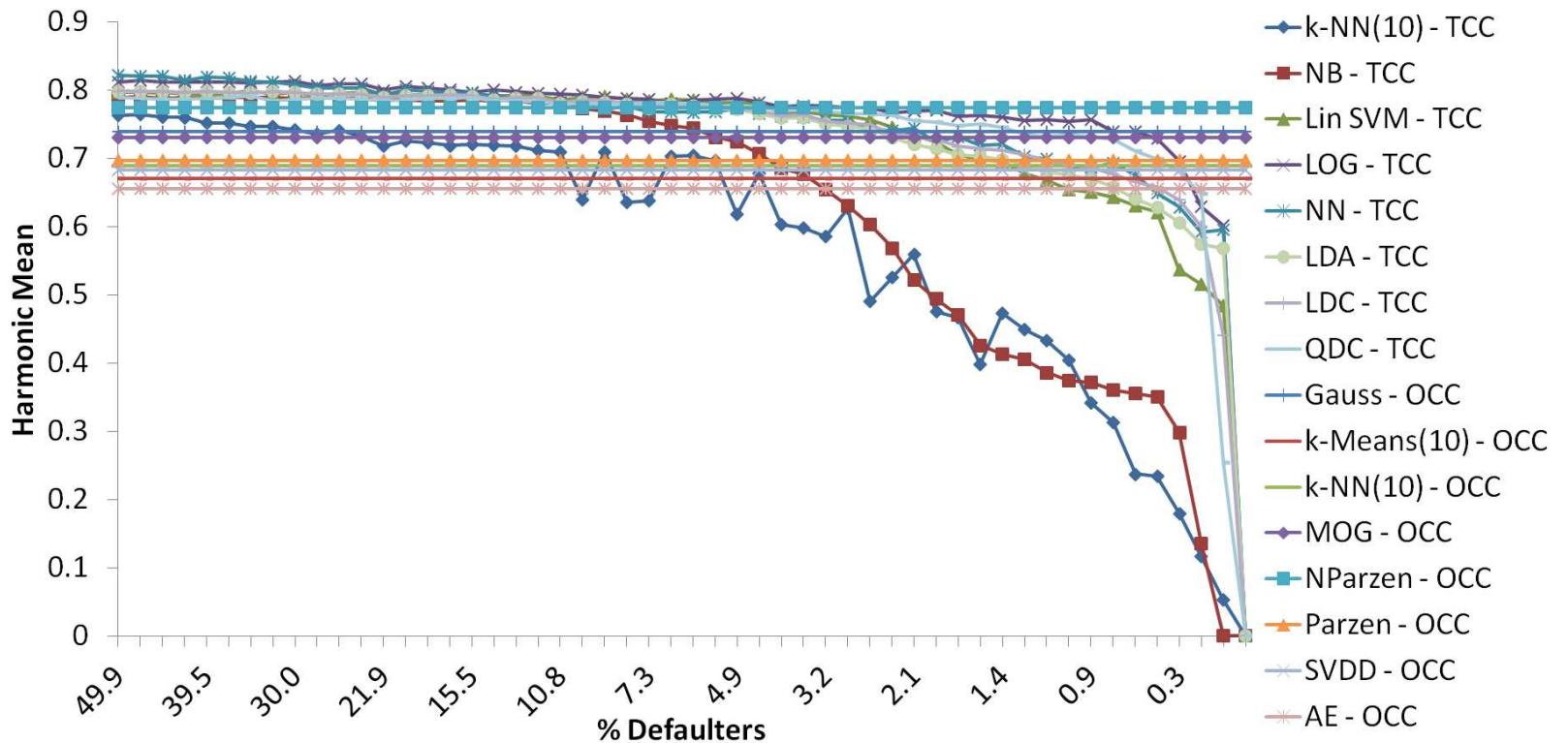


Figure C.29: UCSD: Oversample process and one-class classification process test set harmonic mean performance.

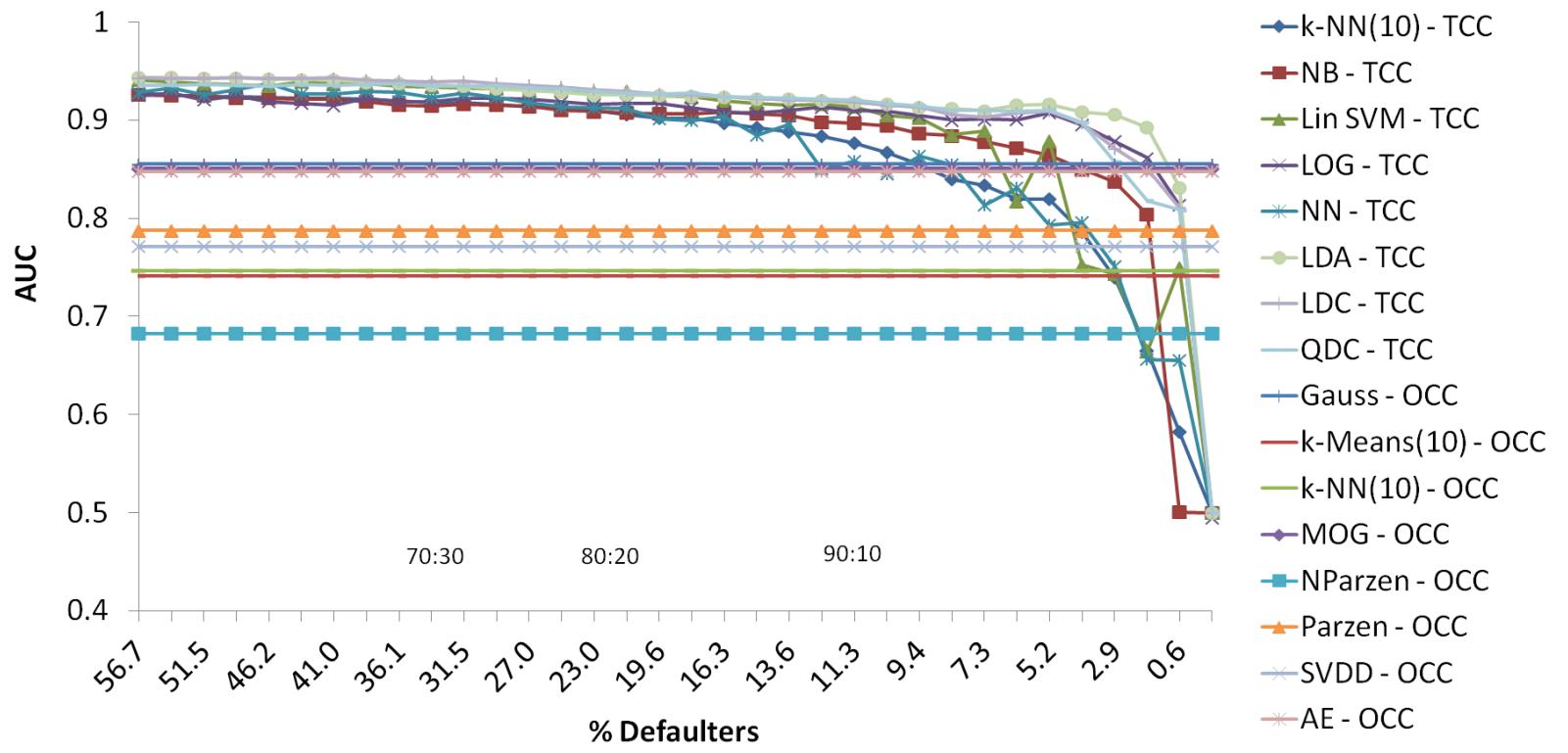


Figure C.30: Australia: Normal process and one-class classification process test set AUC performance.

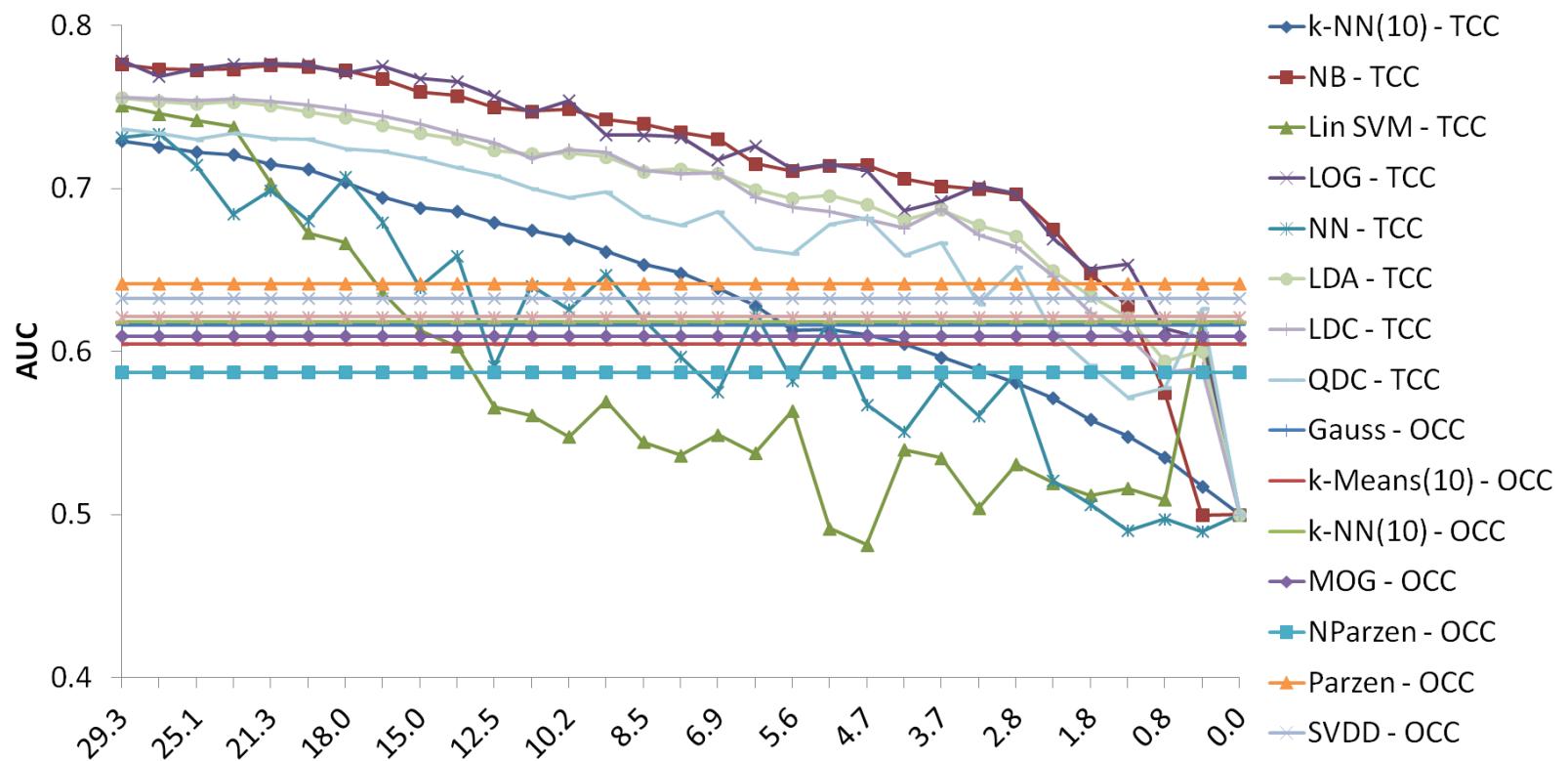


Figure C.31: German: Normal process and one-class classification process test set AUC performance.

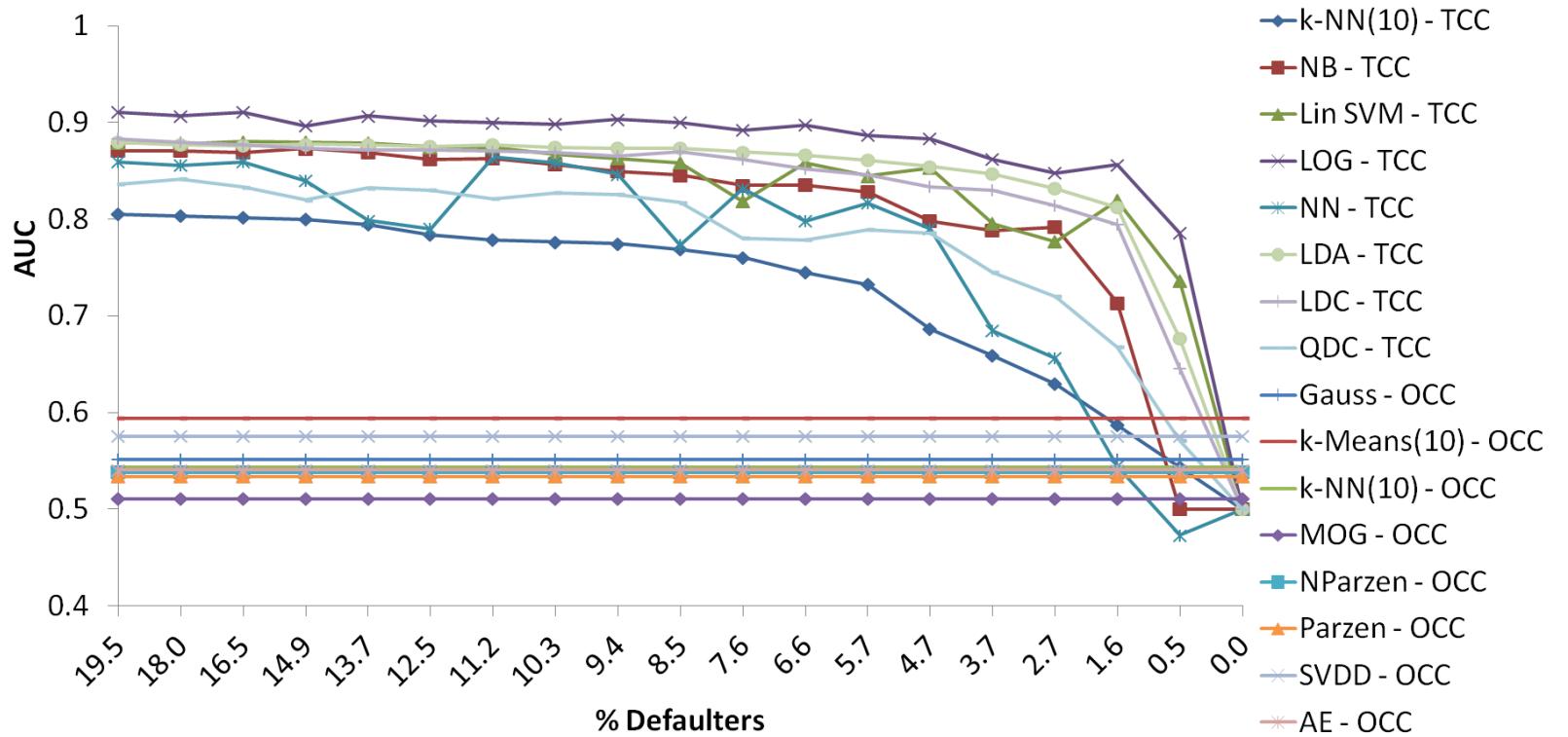


Figure C.32: Iran: Normal process and one-class classification process test set AUC performance.

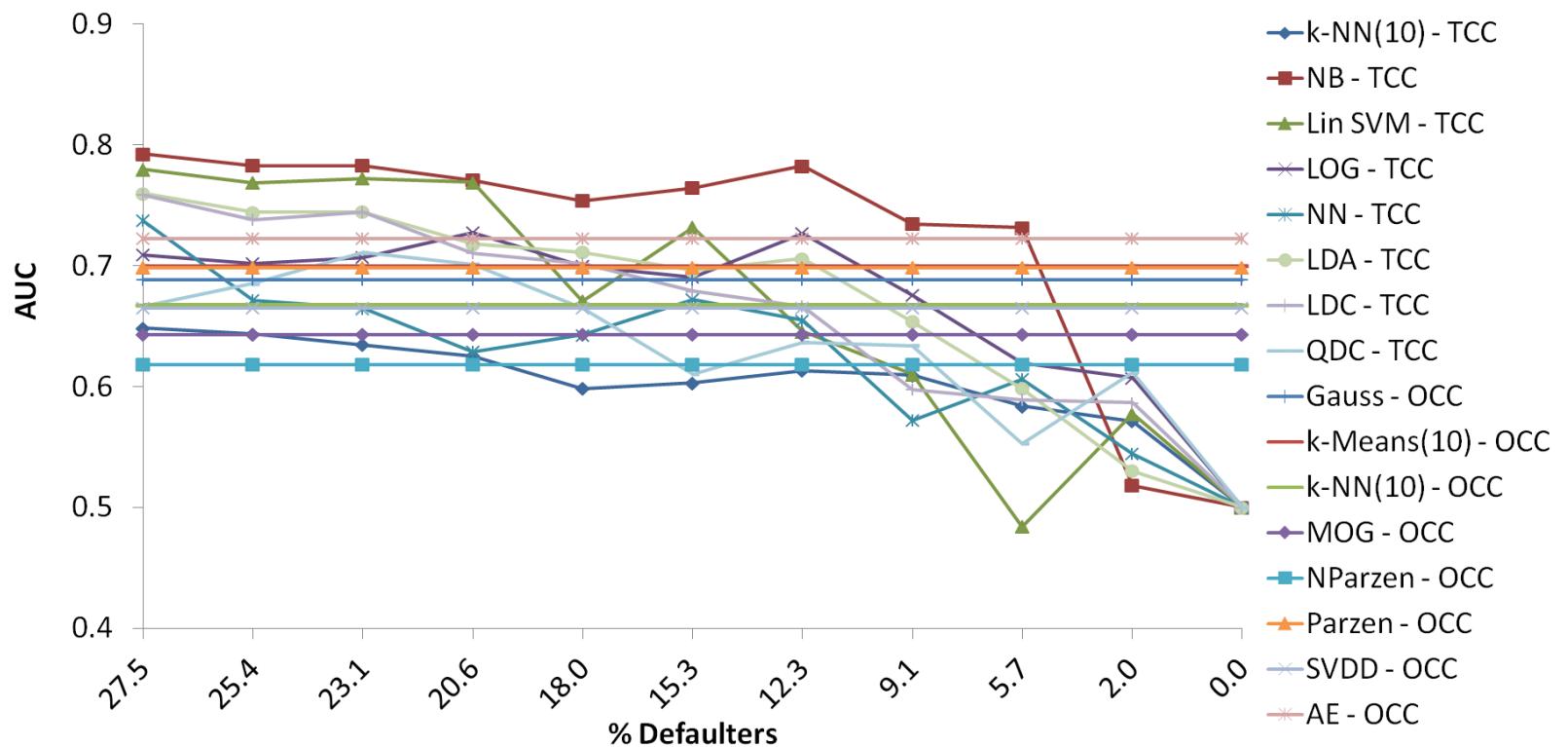


Figure C.33: Japan: Normal process and one-class classification process test set AUC performance.

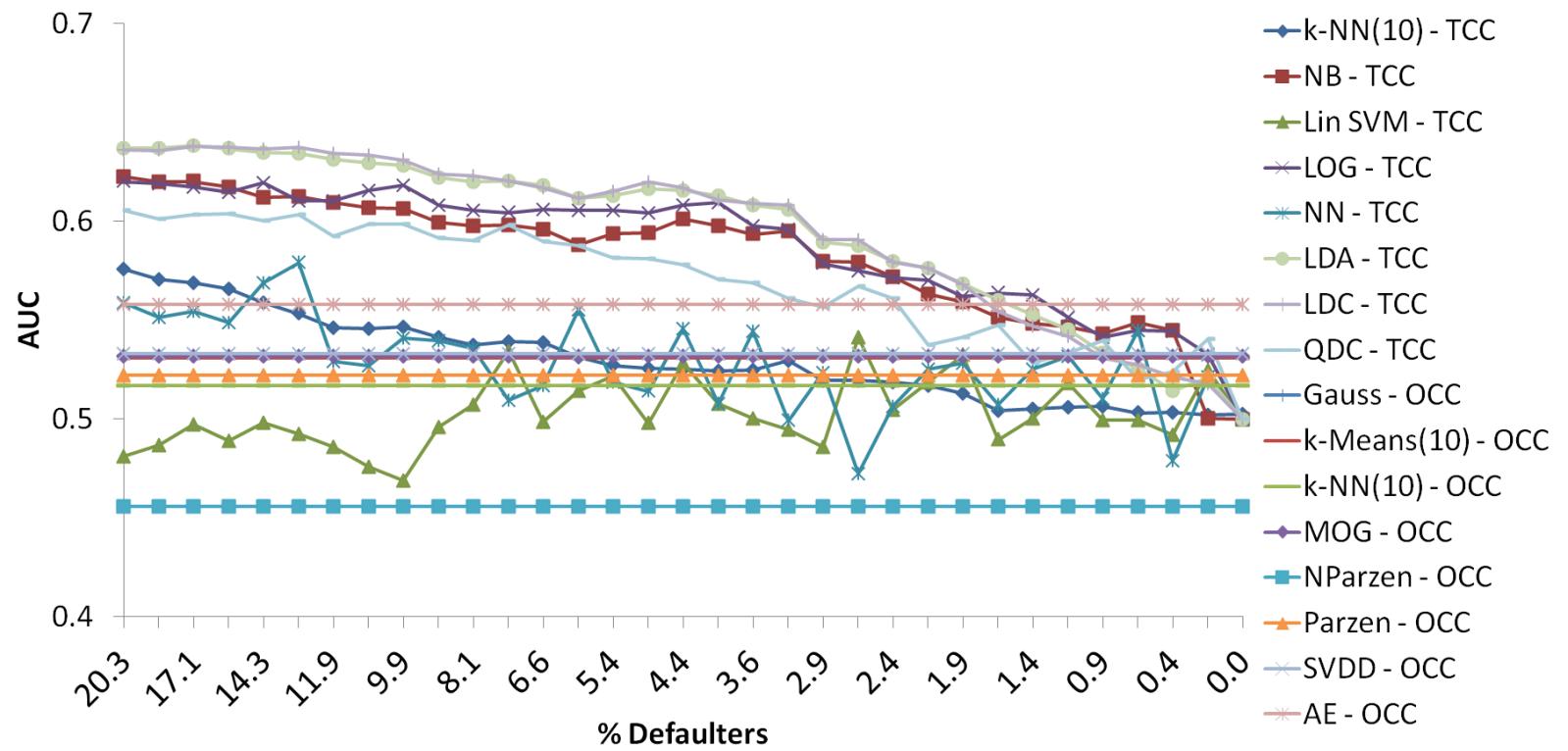


Figure C.34: PAKDD: Normal process and one-class classification process test set AUC performance.

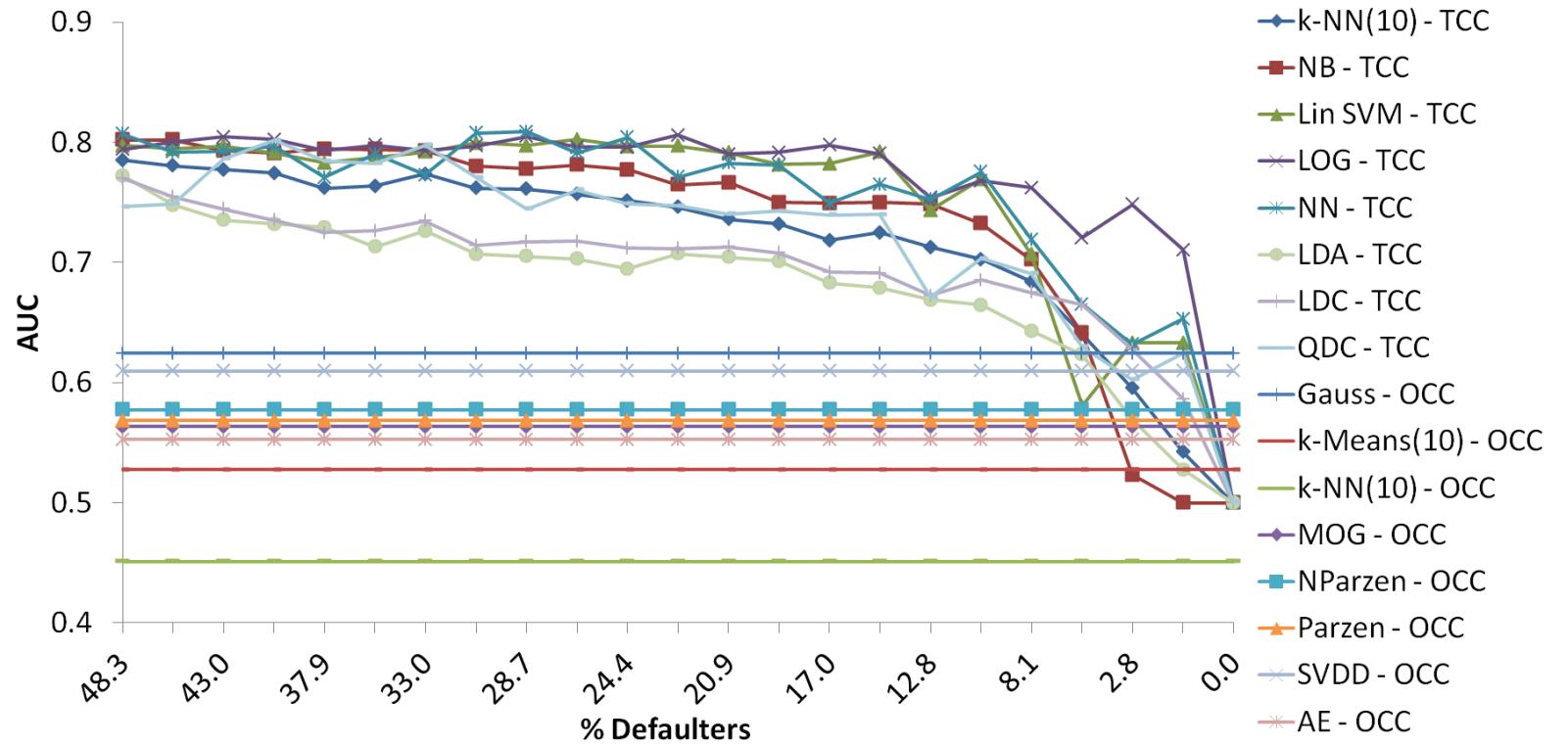


Figure C.35: Poland: Normal process and one-class classification process test set AUC performance.

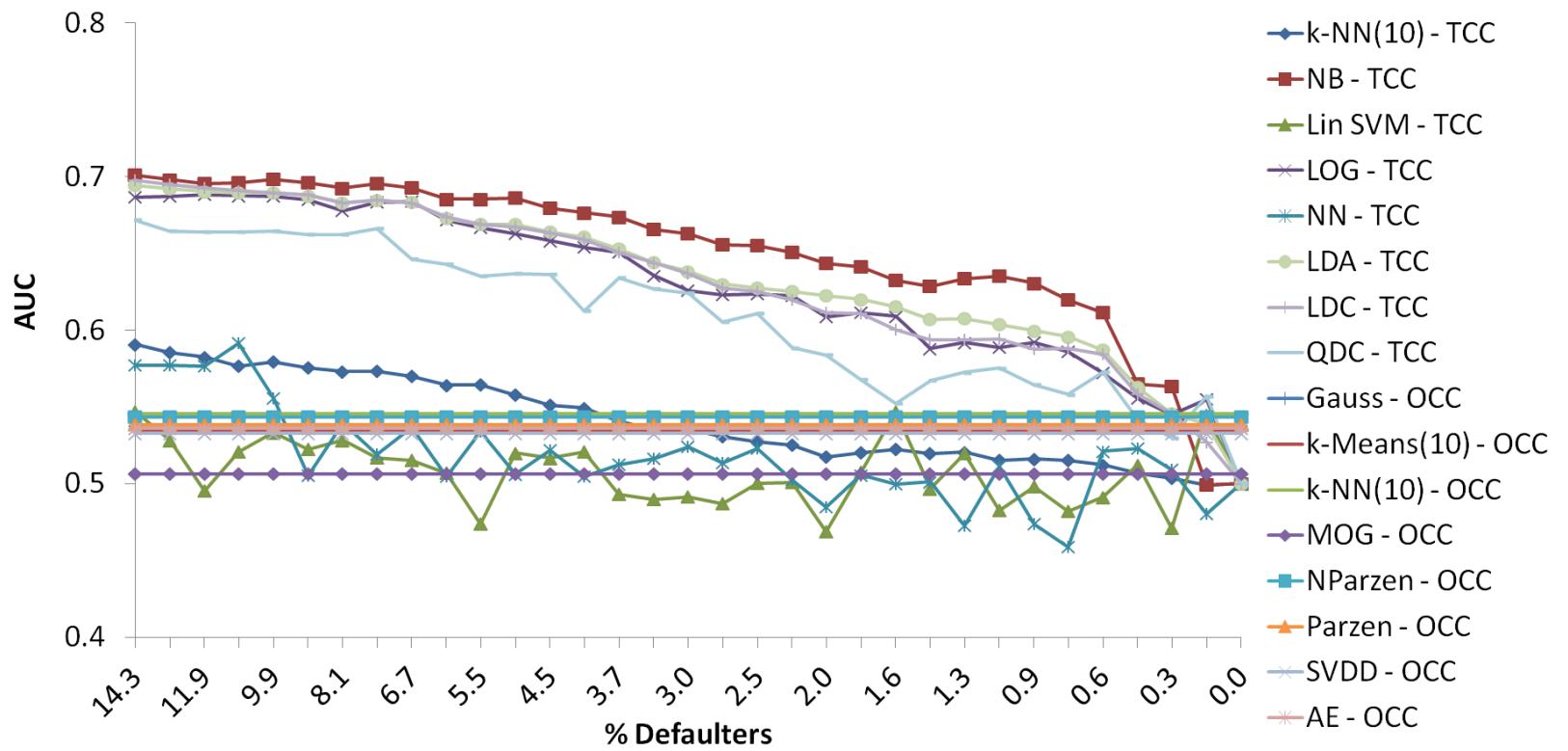


Figure C.36: Spain: Normal process and one-class classification process test set AUC performance.

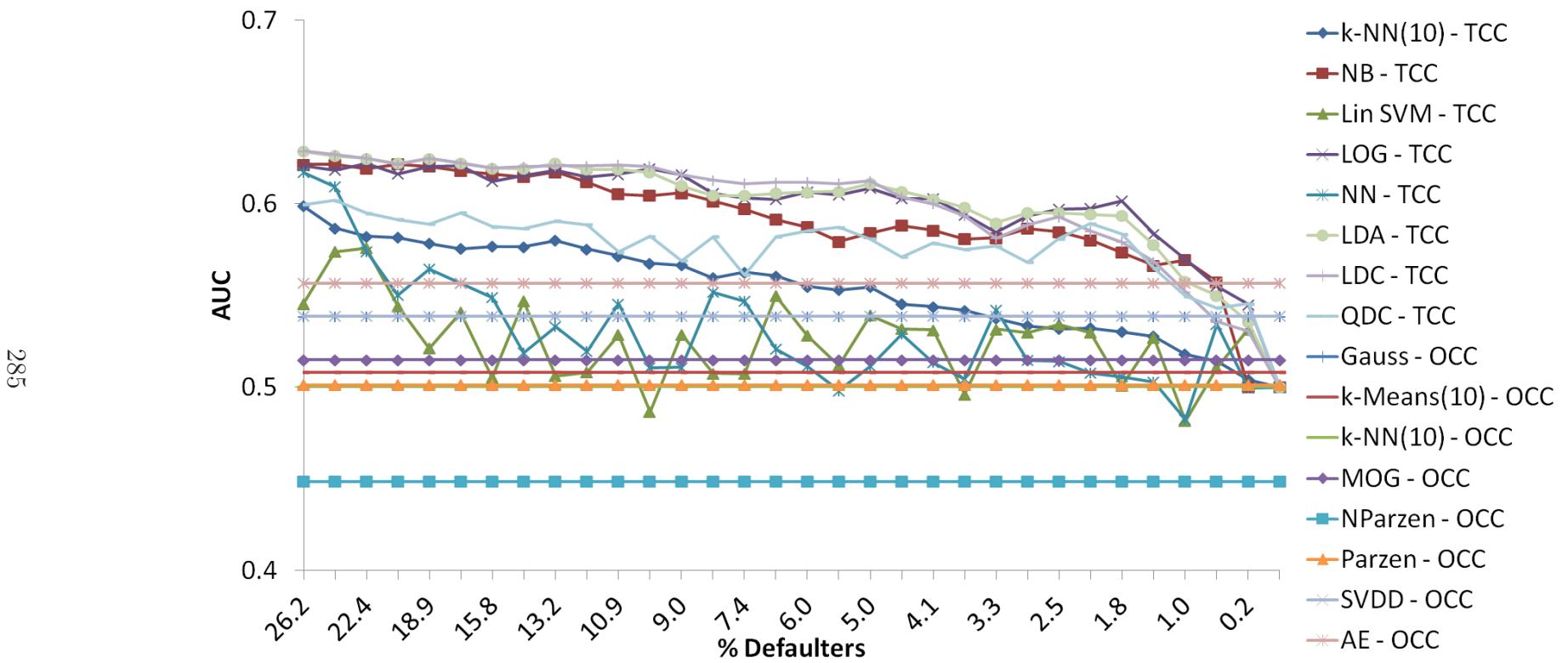


Figure C.37: Thomas: Normal process and one-class classification process test set AUC performance.

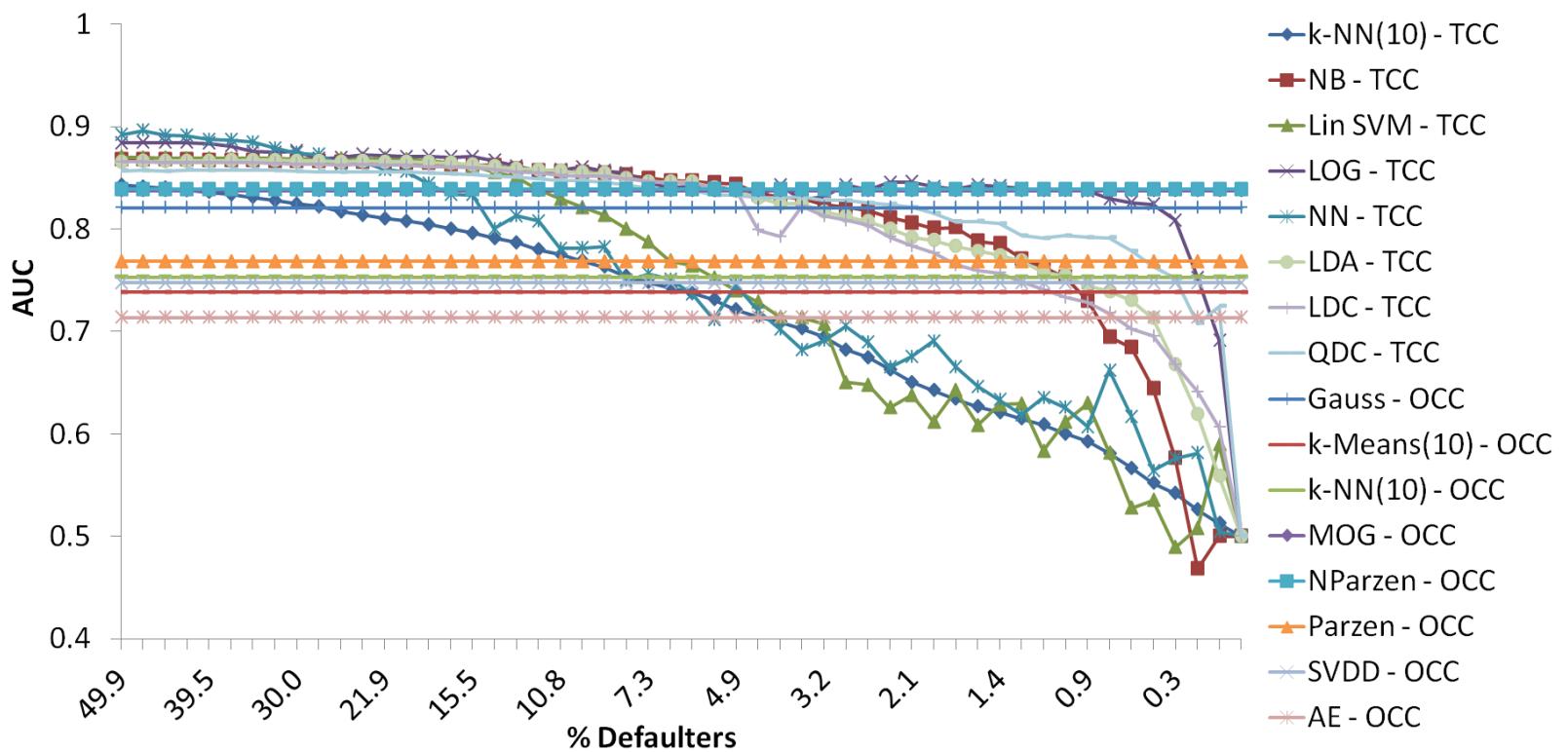


Figure C.38: UCSD: Normal process and one-class classification process test set AUC performance.

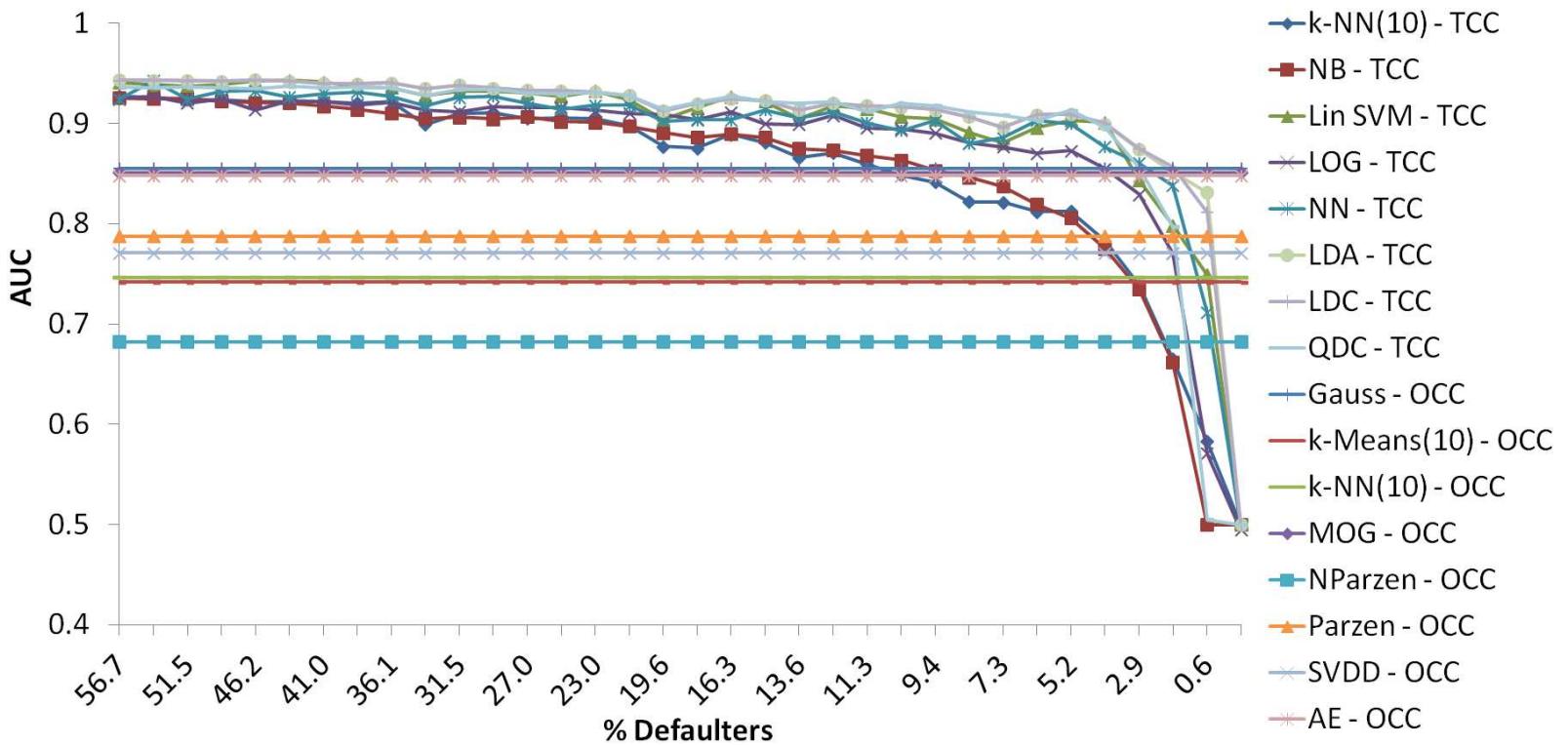


Figure C.39: Australia: Oversample process and one-class classification process test set AUC performance.

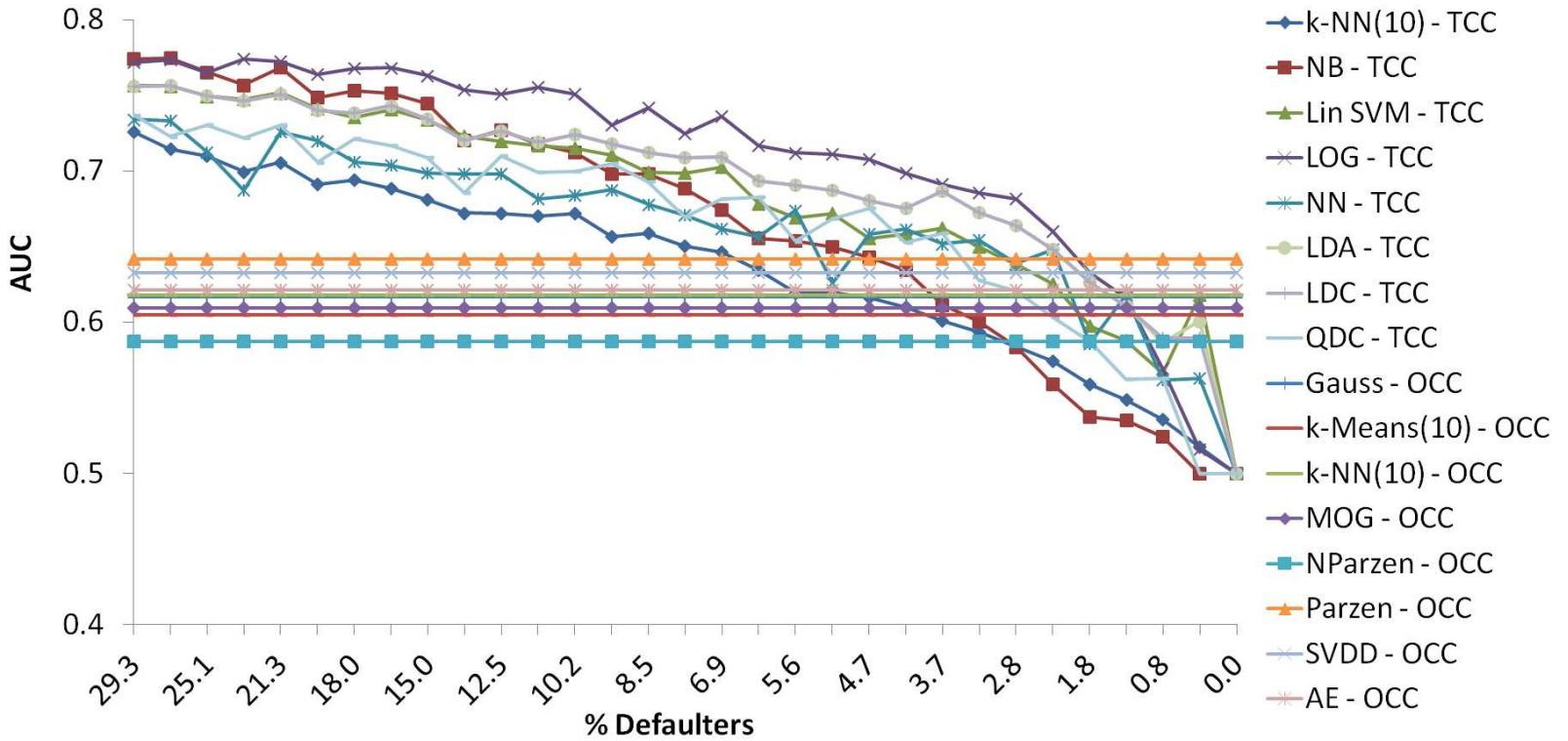


Figure C.40: German: Oversample process and one-class classification process test set AUC performance.

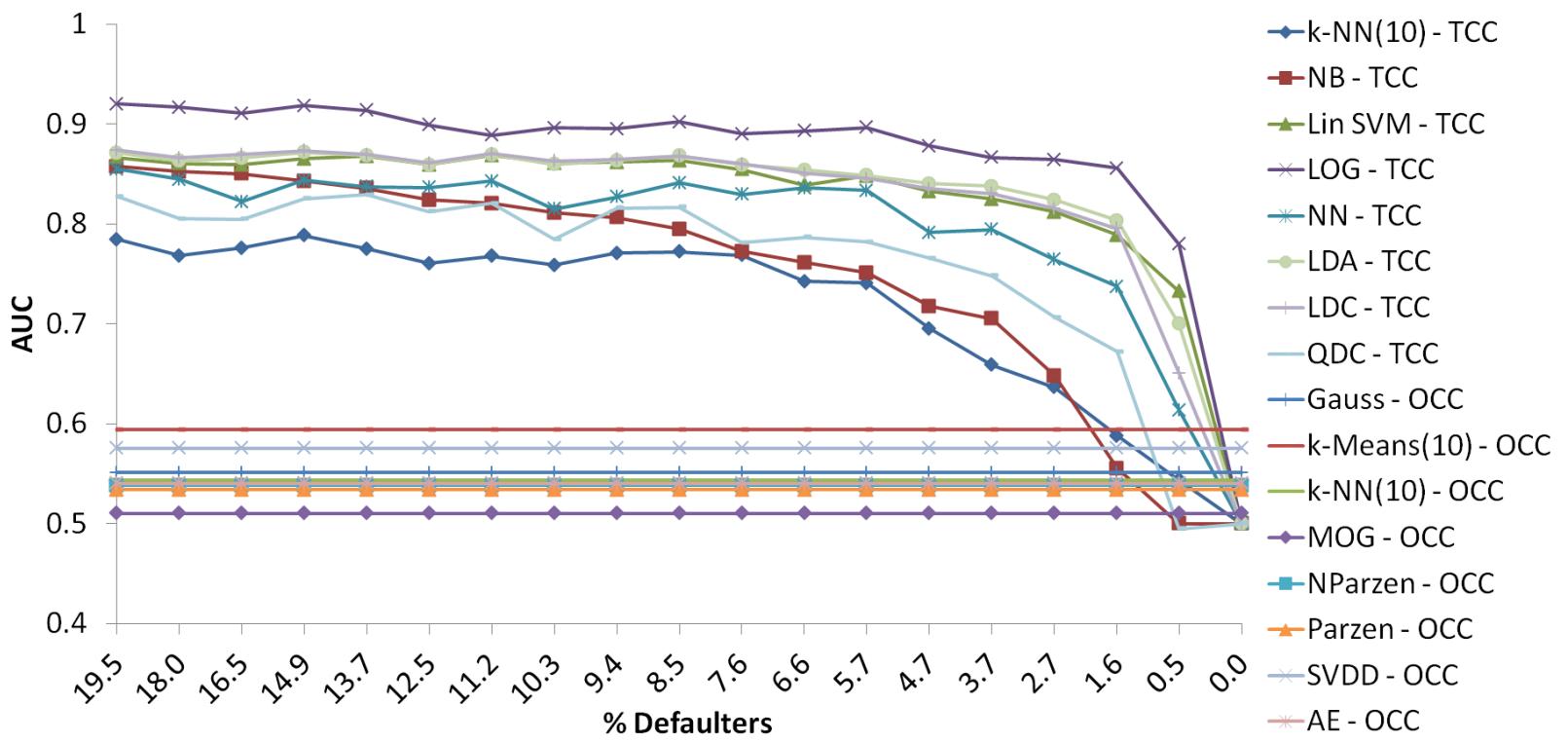


Figure C.41: Iran: Oversample process and one-class classification process test set AUC performance.

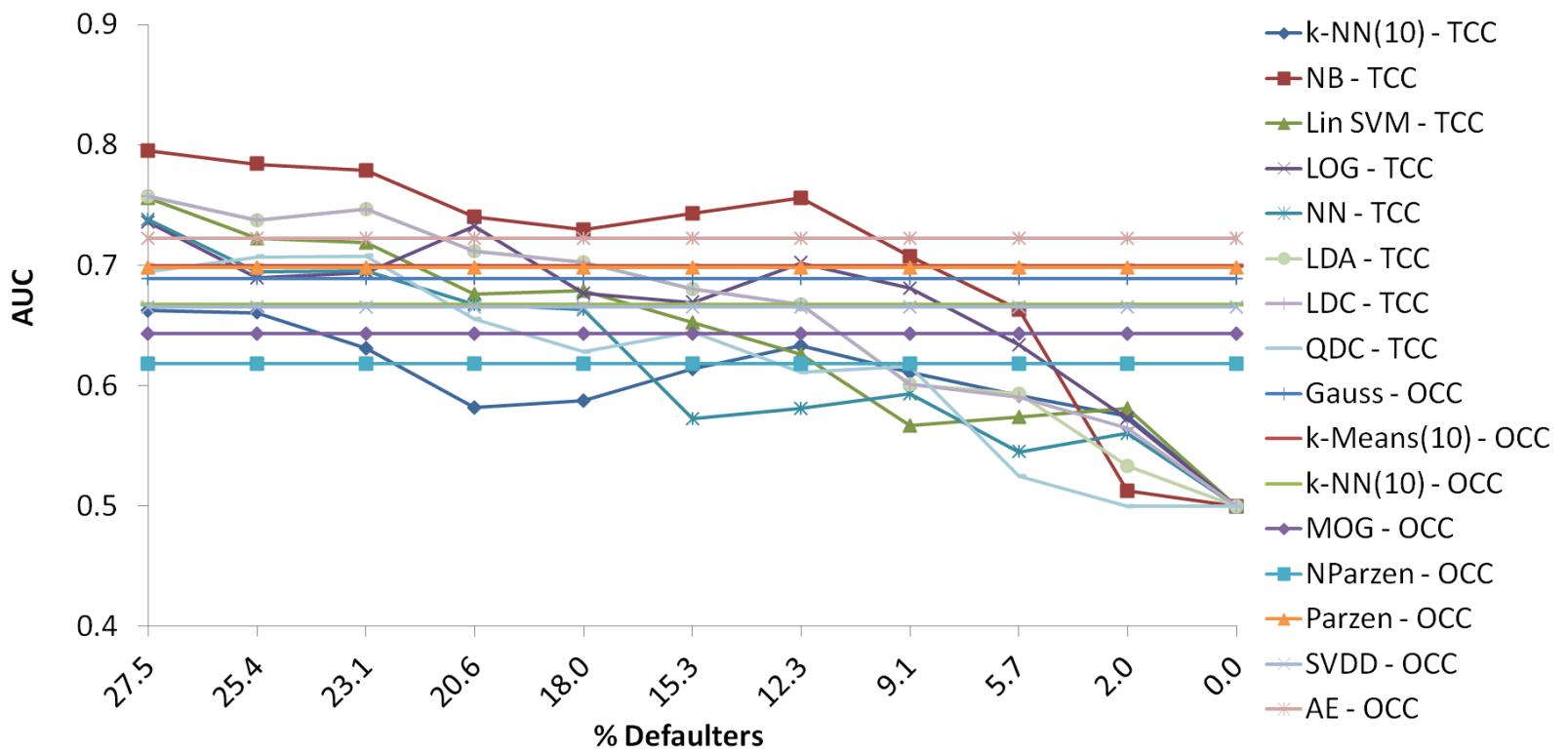


Figure C.42: Japan: Oversample process and one-class classification process test set AUC performance.

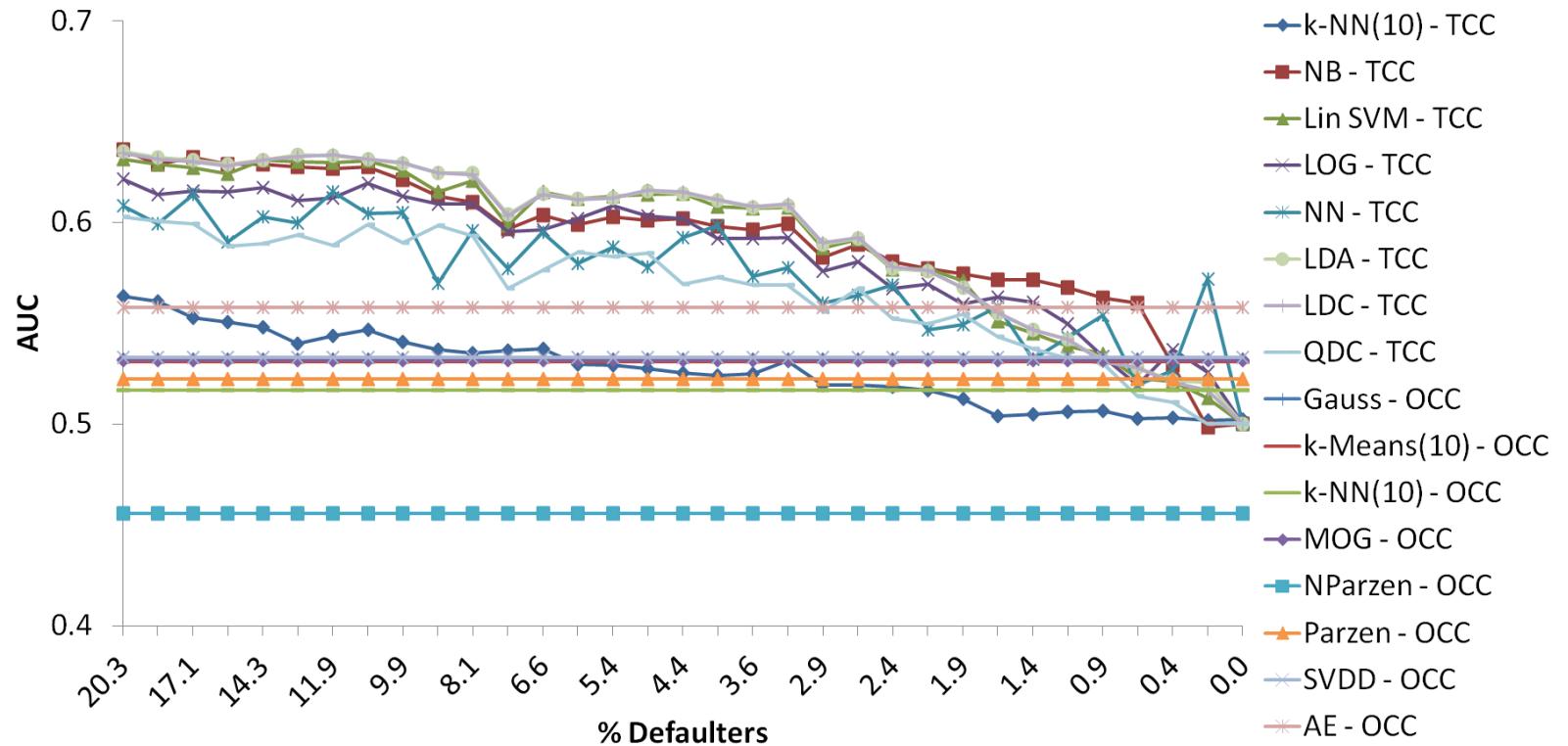


Figure C.43: PAKDD: Oversample process and one-class classification process test set AUC performance.

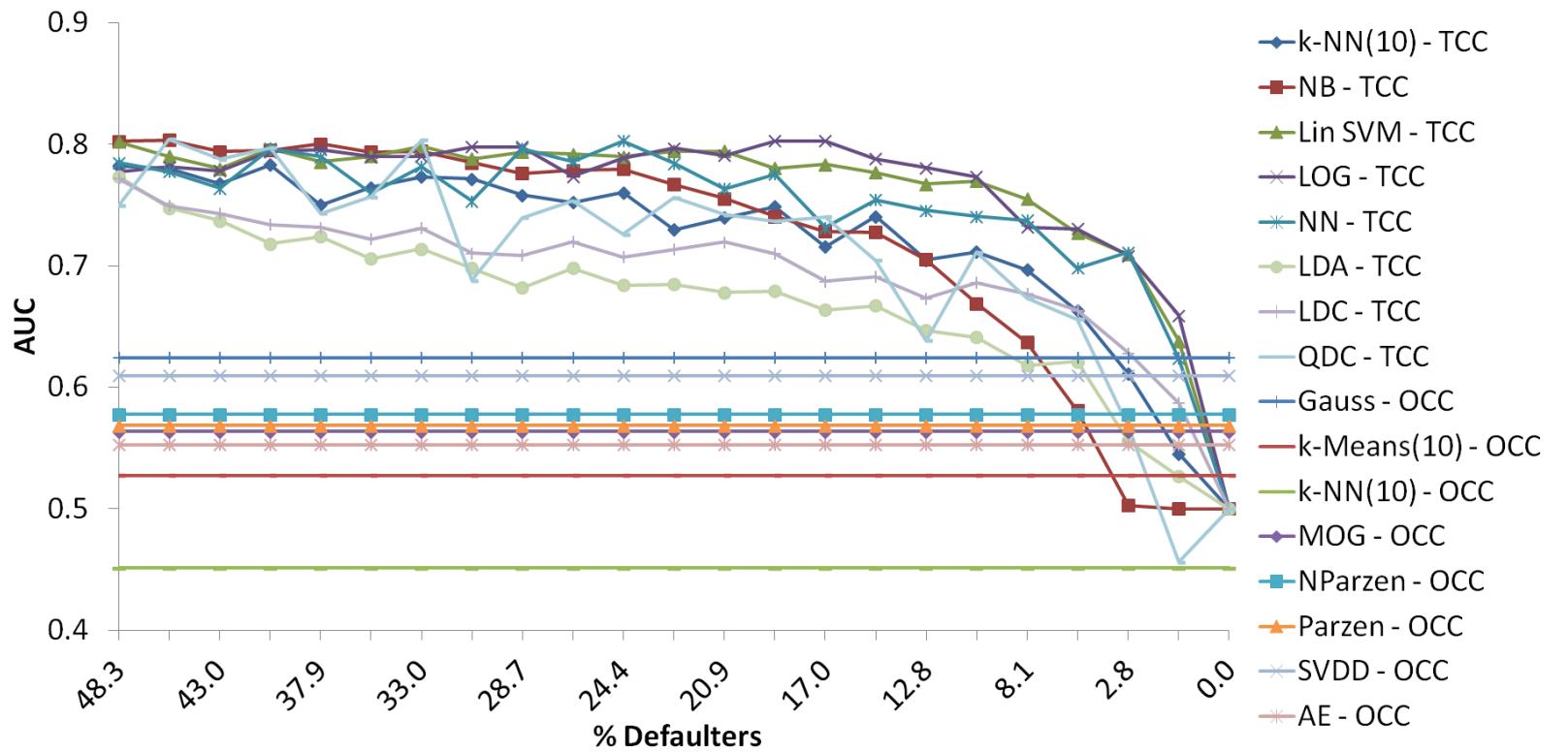


Figure C.44: Poland: Oversample process and one-class classification process test set AUC performance.

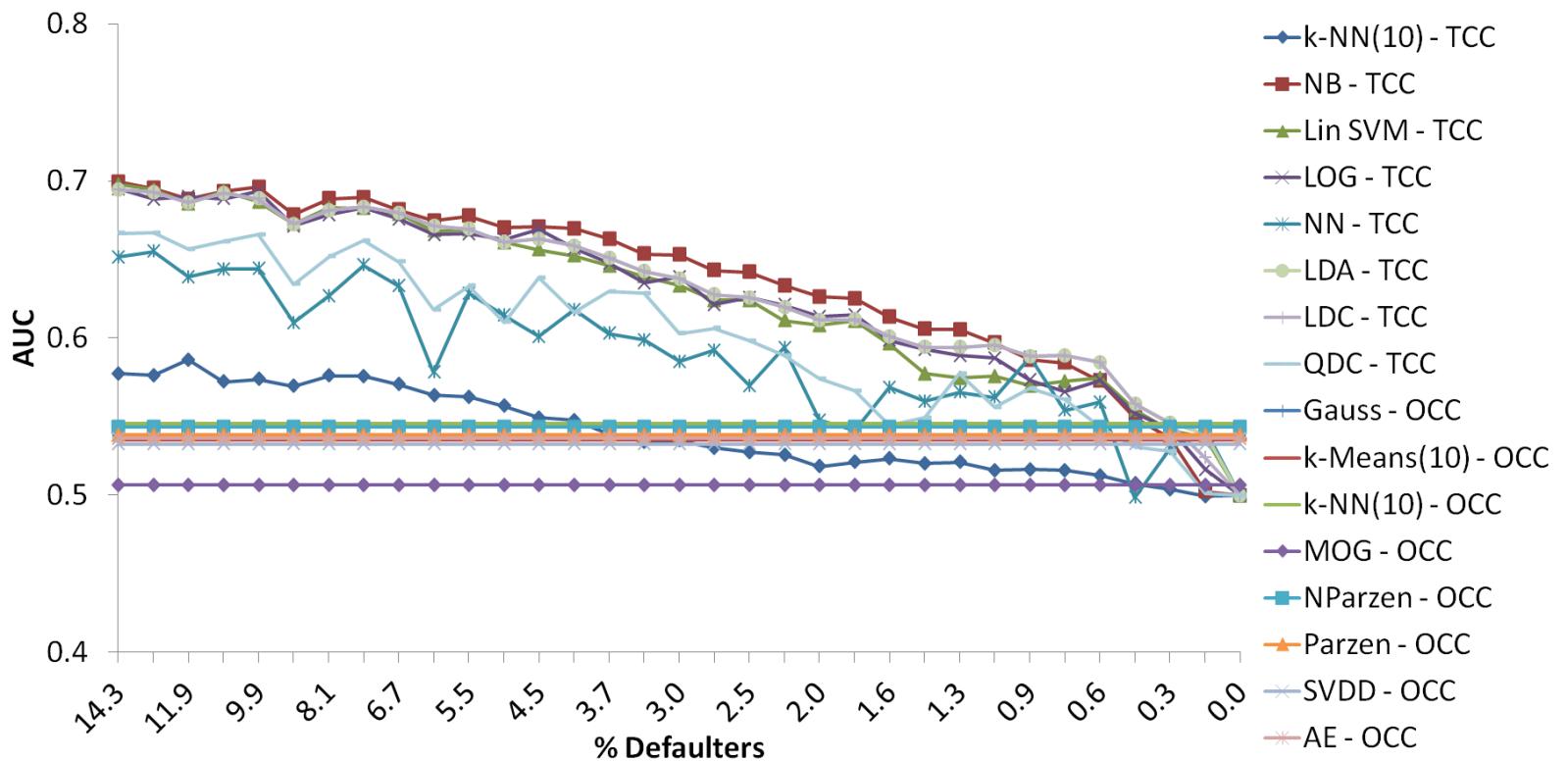


Figure C.45: Spain: Oversample process and one-class classification process test set AUC performance.

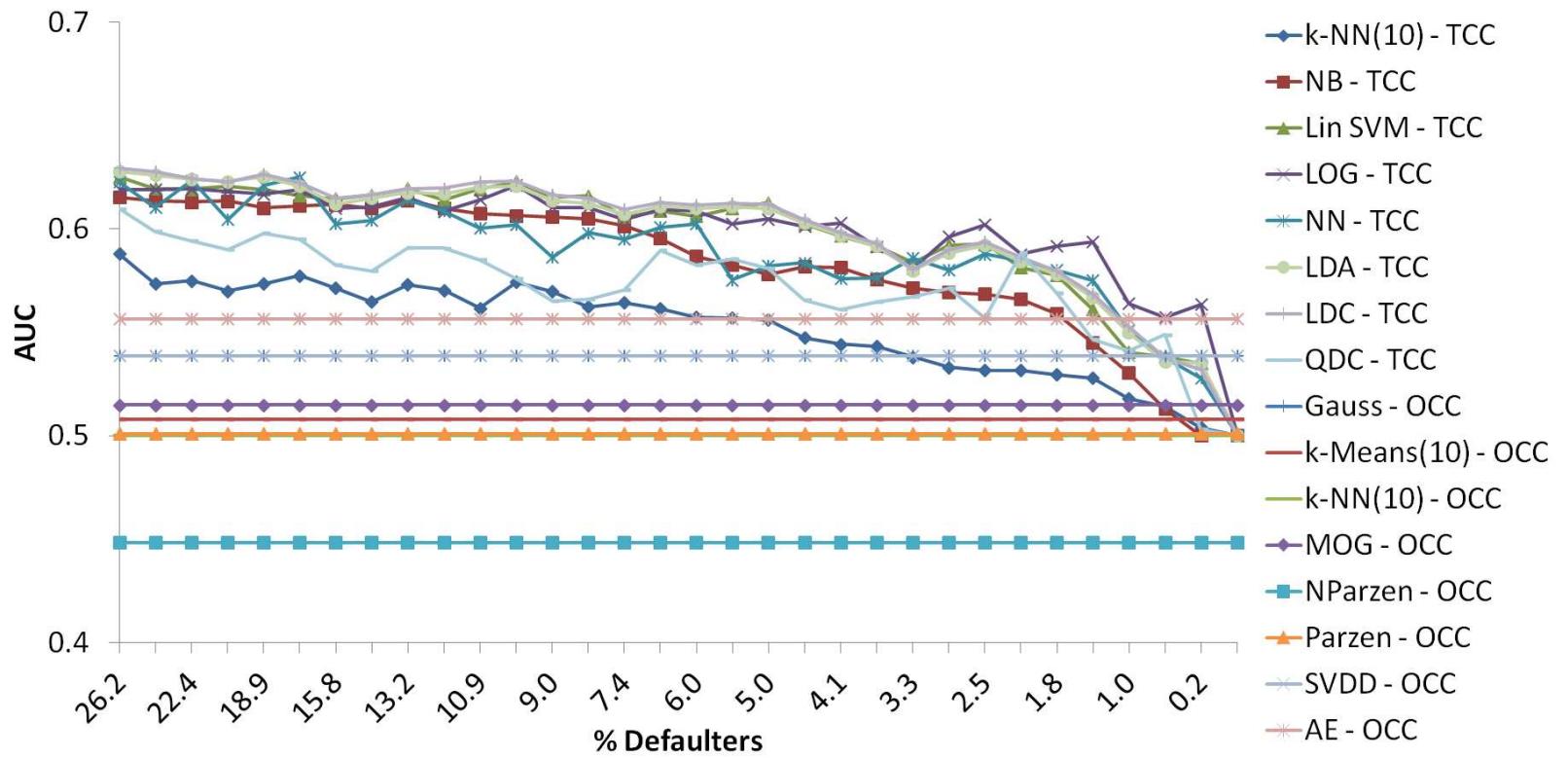


Figure C.46: Thomas: Oversample process and one-class classification process test set AUC performance.

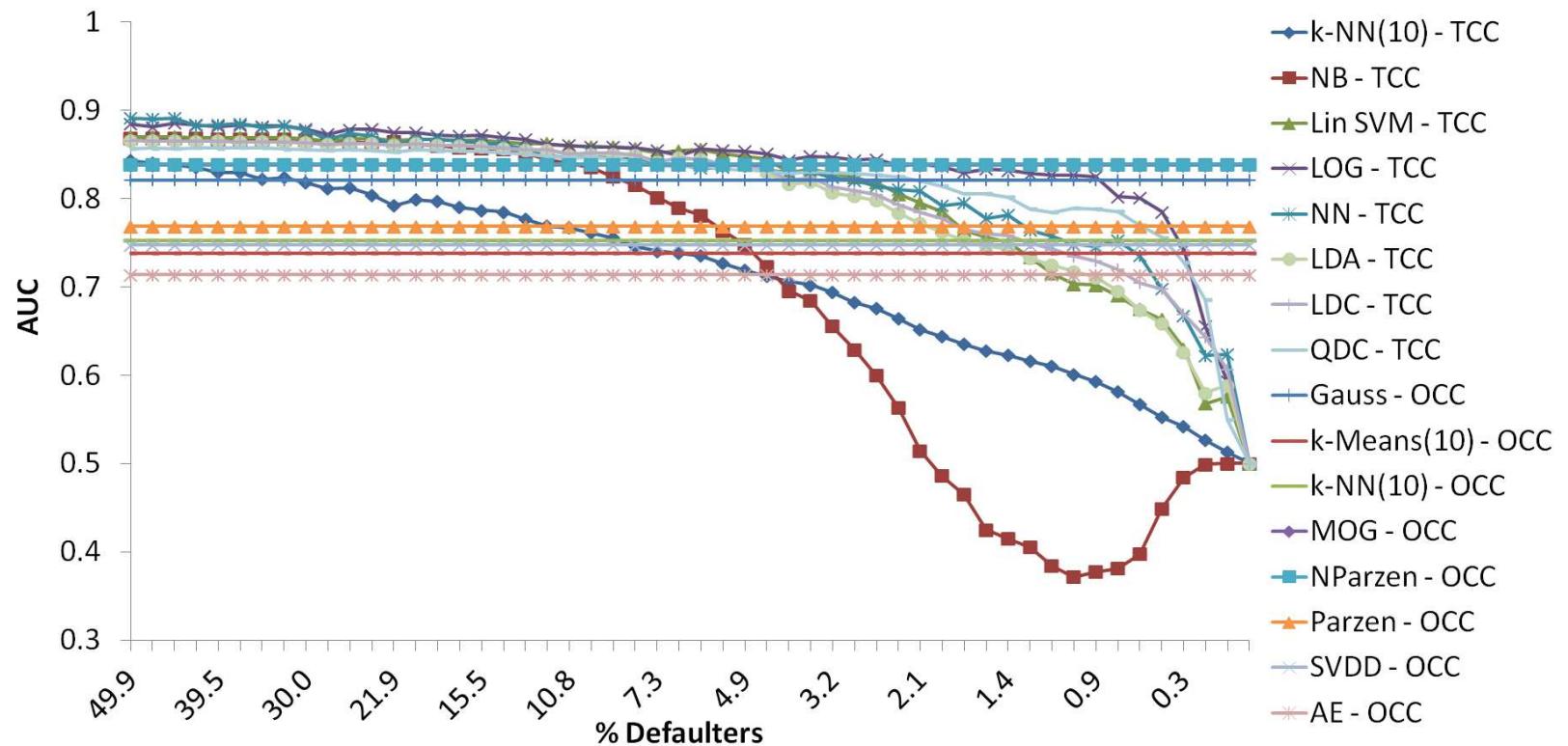


Figure C.47: UCSD: Oversample process and one-class classification process test set AUC performance.

D

APPENDIX

Additional Material for Chapter 7

D.1 Prior Probabilities

This section details the default values of the prior probabilities for the following features: (i) Location; (ii) New Home; and (iii) Loan Rate.

Table D.1: Location prior probabilities.

<i>Location</i>	<i>Prior Probability</i>
Dublin	32.0%
Cork	15.0%
Galway	7.0%
Limerick	4.0%
Waterford	3.0%
Other	39.0%

Table D.2: New Home prior probabilities.

<i>New Home</i>	<i>Prior Probability</i>
New Home	46.0%
Old Home	54.0%

Table D.3: Loan Rate prior probabilities and loan rate values used when calculating the monthly loan repayments are also provided.

<i>Loan Rate</i>	<i>Loan Rate Value</i>	<i>Prior Probability</i>
Tracker Type 1	1.50%	15.50%
Tracker Type 2	2.50%	9.45%
Fixed Type 1	5.35%	45.00%
Fixed Type 2	5.00%	6.70%
Standard Type 1	3.50%	14.00%
Standard Type 2	4.50%	9.35%

D.2 Conditional Prior Probabilities

This section details the default values of the conditional prior probabilities for the following features: (i) Age Group; (ii) Loan-to-Value; (iii) First-Time-Buyer; (iv) Loan Value Group; (v) Income Group; (vi) Loan Term; (vii) Occupation; (viii) Employment; (ix) Household; (x) Education; (xi) Expenses-to-Household; (xii) Expenses-to-Income.

Table D.4: Age group conditional prior probabilities. Each column should total 100%.

<i>Age</i>	<i>FTB</i>	<i>Not FTB</i>
18 - 25	18.0%	4.0%
26 - 30	40.0%	16.0%
31 - 35	23.0%	23.0%
36 - 40	10.0%	20.0%
41 - 45	5.0%	15.0%
46 - 55	4.0%	22.0%

Table D.5: LTV conditional prior probabilities. NH = New Home, OH = Old Home, NFTB = Not First-Time-buyer. Each row should total 100%.

<i>LTV</i>	0.45	0.55	0.6	0.65	0.7	0.75	0.85	0.93	0.975	1
FTB & NH & Dublin	1%	2%	2%	3%	4%	8%	13%	27%	20%	20%
FTB & OH & Dublin	1%	2%	2%	3%	4%	7%	16%	43%	5%	17%
NFTB & NH & Dublin	2%	3%	4%	11%	15%	17%	23%	18%	2%	5%
NFTB & OH & Dublin	3%	3%	10%	12%	13%	15%	18%	19%	2%	5%
FTB & NH & Not Dublin	3%	3%	3%	4%	11%	8%	11%	20%	9%	28%
FTB & OH & Not Dublin	1%	1%	2%	3%	4%	7%	16%	42%	4%	20%
NFTB & NH & Not Dublin	4%	8%	12%	10%	12%	15%	16%	13%	1%	9%
NFTB & OH & Not Dublin	3%	6%	9%	16%	15%	14%	17%	14%	1%	5%

Table D.6: First-Time-Buyer (FTB) conditional prior probability (CPP). NFTB = Not First-Time-buyer.

<i>FTB</i>	<i>CPP</i>
FTB & New Home & Dublin	41.0%
NFTB & New Home & Dublin	59.0%
FTB & Old Home & Dublin	30.0%
NFTB & Old Home & Dublin	70.0%
FTB & New Home & Cork	38.0%
NFTB & New Home & Cork	62.0%
FTB & Old Home & Cork	30.0%
NFTB & Old Home & Cork	70.0%
FTB & New Home & Galway	38.0%
NFTB & New Home & Galway	62.0%
FTB & Old Home & Galway	30.0%
NFTB & Old Home & Galway	70.0%
FTB & New Home & Limerick	38.0%
NFTB & New Home & Limerick	62.0%
FTB & Old Home & Limerick	30.0%
NFTB & Old Home & Limerick	70.0%
FTB & New Home & Waterford	38.0%
NFTB & New Home & Waterford	62.0%
FTB & Old Home & Waterford	30.0%
FTB & Old Home & Waterford	70.0%
FTB & New Home & Other	38.0%
NFTB & New Home & Other	62.0%
FTB & Old Home & Other	30.0%
NFTB & Old Home & Other	70.0%

Table D.7: Loan Value conditional prior probabilities. NH = New Home, OH = Old Home, NFTB = Not First-Time-buyer. Each row should total 100%.

<i>Loan Value</i>	50k- 100k	100k- 150k	150k- 200k	200k- 250k	250k- 300k	300k- 350k	350k- 400k	400k- 450k	450k- 900k
FTB & NH & Dublin	0.8%	2.9%	13.4%	30.0%	27.0%	12.2%	7.0%	6.0%	0.7%
NFTB & NH & Dublin	2.0%	4.0%	10.0%	15.0%	18.0%	13.0%	14.0%	16.0%	8.0%
FTB & OH & Dublin	1.3%	2.5%	6.2%	13.0%	25.0%	28.0%	10.0%	11.0%	3.0%
NFTB & OH & Dublin	8.0%	5.0%	8.0%	12.0%	15.3%	12.7%	12.0%	15.0%	12.0%
FTB & NH & Cork/Galway	5.5%	14.3%	31.4%	23.8%	16.5%	7.0%	1.0%	0.5%	0.0%
NFTB & NH & Cork/Galway	8.0%	12.5%	21.2%	22.2%	14.6%	10.5%	4.0%	5.0%	2.0%
FTB & OH & Cork/Galway	4.2%	8.3%	18.4%	27.0%	22.5%	12.0%	5.0%	2.3%	0.3%
NFTB & OH & Cork/Galway	10.9%	13.3%	19.3%	16.5%	11.0%	16.0%	9.0%	2.0%	2.0%
FTB & NH & Limerick/Waterford	5.5%	16.2%	34.4%	24.9%	12.5%	5.0%	1.0%	0.5%	0.0%
NFTB & NH & Limerick/Waterford	10.0%	17.5%	23.0%	23.0%	15.0%	5.0%	4.0%	2.0%	0.5%
FTB & OH & Limerick/Waterford	5.7%	12.3%	20.4%	25.0%	21.5%	11.0%	2.0%	1.3%	0.8%
NFTB & OH & Limerick/Waterford	23.0%	19.0%	22.3%	17.5%	10.5%	2.0%	2.3%	1.9%	1.5%
FTB & NH & Other	4.5%	15.2%	35.5%	28.0%	11.1%	4.2%	1.0%	0.5%	0.0%
NFTB & NH & Other	9.5%	10.9%	23.5%	17.0%	18.2%	8.0%	5.0%	5.0%	2.9%
FTB & OH & Other	5.2%	14.0%	23.5%	27.8%	17.8%	8.0%	2.0%	1.0%	0.7%
NFTB & OH & Other	15.5%	20.3%	21.4%	17.4%	13.9%	4.0%	2.5%	3.0%	2.0%

Table D.8: Income Group conditional prior probabilities. Each row should total 100%.

<i>Income</i>	40k- 60k	60k- 80k	80k- 100k	100k- 120k	120k- 150k	150k+
FTB & Dublin	6.6%	14.8%	19.9%	19.2%	20.0%	19.5%
NFTB & Dublin	4.0%	6.2%	9.6%	11.1%	30.0%	39.1%
FTB & Not Dublin	17.3%	21.5%	21.3%	15.7%	12.0%	12.2%
NFTB & Not Dublin	9.0%	11.1%	13.1%	12.9%	20.0%	33.9%

Table D.9: Loan Term conditional prior probabilities. Each row should total 100%.

<i>Years</i>	20	25	30	35	40
FTB & Dublin & 18-25	1%	3%	8%	80%	8%
FTB & Dublin & 26-30	2%	6%	16%	72%	4%
FTB & Dublin & 31-35	2%	6%	16%	72%	4%
FTB & Dublin & 36-40	2%	6%	16%	72%	4%
FTB & Dublin & 41-45	10%	73%	12%	5%	0%
FTB & Dublin & 46-55	12%	75%	8%	5%	0%
Not FTB & Dublin & 18-25	12%	14%	21%	48%	5%
Not FTB & Dublin & 26-30	23%	28%	22%	24%	3%
Not FTB & Dublin & 31-35	22%	32%	19%	24%	3%
Not FTB & Dublin & 36-40	22%	32%	19%	24%	3%
Not FTB & Dublin & 41-45	22%	32%	25%	20%	1%
Not FTB & Dublin & 46-55	22%	32%	24%	22%	0%
FTB & Not Dublin & 18-25	3%	5%	17%	68%	7%
FTB & Not Dublin & 26-30	5%	9%	18%	62%	6%
FTB & Not Dublin & 31-35	6%	11%	20%	59%	4%
FTB & Not Dublin & 36-40	6%	11%	20%	59%	4%
FTB & Not Dublin & 41-45	14%	29%	30%	25%	2%
FTB & Not Dublin & 46-55	34%	39%	14%	13%	0%
Not FTB & Not Dublin & 18-25	6%	9%	12%	41%	32%
Not FTB & Not Dublin & 26-30	14%	15%	25%	32%	14%
Not FTB & Not Dublin & 31-35	22%	33%	24%	19%	2%
Not FTB & Not Dublin & 36-40	22%	33%	24%	19%	2%
Not FTB & Not Dublin & 41-45	13%	46%	29%	10%	2%
Not FTB & Not Dublin & 46-55	14%	68%	10%	8%	0%

Table D.10: Occupation conditional prior probabilities. M/E = Managerial/Employer. The column of each division should total 100%.

<i>Income</i>	40k- 60k	60k- 80k	80k- 100k	100k- 120k	120k- 150k	150k+
FTB & Dublin & M/E	8%	13%	13%	35%	40%	48%
FTB & Dublin & Office:Salaried	34%	41%	42%	42%	42%	39%
FTB & Dublin & Skilled	35%	28%	27%	10%	7%	3%
FTB & Dublin & Semi-Skilled	12%	7%	7%	3%	1%	0%
FTB & Dublin & Manual	1%	1%	1%	0%	0%	0%
FTB & Dublin & Self-Employed	10%	10%	10%	10%	10%	10%
Not FTB & Dublin & M/E	12%	26%	33%	51%	66%	78%
Not FTB & Dublin & Office:Salaried	48%	48%	45%	32%	22%	11%
Not FTB & Dublin & Skilled	22%	12%	9%	7%	2%	1%
Not FTB & Dublin & Semi-Skilled	8%	4%	3%	0%	0%	0%
Not FTB & Dublin & Manual	0%	0%	0%	0%	0%	0%
Not FTB & Dublin & Self-Employed	10%	10%	10%	10%	10%	10%
FTB & Not Dublin & M/E	3%	5%	9%	14%	18%	20%
FTB & Not Dublin & Office:Salaried	17%	15%	16%	26%	39%	40%
FTB & Not Dublin & Skilled	40%	35%	35%	30%	25%	25%
FTB & Not Dublin & Semi-Skilled	20%	25%	20%	10%	3%	0%
FTB & Not Dublin & Manual	10%	10%	10%	10%	5%	5%
FTB & Not Dublin & Self-Employed	10%	10%	10%	10%	10%	10%
Not FTB & Not Dublin & M/E	10%	24%	33%	50%	63%	73%
Not FTB & Not Dublin & Office:Salaried	42%	35%	32%	24%	18%	9%
Not FTB & Not Dublin & Skilled	28%	22%	16%	9%	4%	3%
Not FTB & Not Dublin & Semi-Skilled	7%	6%	6%	4%	0%	0%
Not FTB & Not Dublin & Manual	3%	3%	3%	3%	5%	5%
Not FTB & Not Dublin & Self-Employed	10%	10%	10%	10%	10%	10%

Table D.11: Employment conditional prior probabilities. M/E = Managerial/Employer. Each column should total 100%.

<i>Occupation</i>	<i>M/E</i>	<i>Office:Salaried</i>	<i>Skilled</i>	<i>Semi-Skilled</i>	<i>Manual</i>	<i>Self-Employed</i>
Agriculture	1.0%	1.0%	5.0%	5.0%	75.0%	5.20%
Construction	0.5%	2.0%	30.0%	40.0%	0.0%	12.59%
Wholesale/Retail	12.0%	25.0%	10.0%	3.0%	3.0%	14.26%
Transportation/Storage	2.5%	2.0%	3.0%	10.0%	3.0%	4.43%
Hospitality	10.5%	6.0%	5.0%	5.0%	0.0%	6.26%
Information/Communication	8.0%	8.0%	0.0%	0.0%	0.0%	3.22%
Professional/Scientific/Technical	15.0%	7.0%	0.0%	0.0%	0.0%	5.23%
Admin/Support services	6.0%	7.0%	3.0%	0.0%	0.0%	3.71%
Public administration	5.0%	5.0%	5.0%	5.0%	5.0%	4.88%
Education	11.5%	10.0%	0.0%	0.0%	0.0%	6.55%
Health	13.0%	10.0%	5.0%	3.0%	0.0%	10.11%
Industry	2.0%	1.0%	30.0%	25.0%	10.0%	14.01%
Financial	9.0%	12.0%	0.0%	0.0%	0.0%	4.82%
Other	4.0%	4.0%	4.0%	4.0%	4.0%	4.74%

Table D.12: Household conditional prior probabilities. The column of each division should total 100%.

<i>Income</i>	40k- 60k	60k- 80k	80k- 100k	100k- 120k	120k- 150k	150k+
FTB & 1 Adult, No Child < 18	30%	30%	25%	20%	22%	18%
FTB & 1 Adult, 1+ Child < 18	33%	20%	10%	5%	1%	2%
FTB & 2 Adults, No Child < 18	13%	17%	22%	30%	38%	33%
FTB & 3+ adults, No Child < 18	2%	6%	8%	8%	2%	3%
FTB & 2 Adults, 1+ Child < 18	12%	17%	25%	27%	27%	34%
FTB & Other	10%	10%	10%	10%	10%	10%
Not FTB & 1 Adult, No Child < 18	25%	27%	17%	12%	10%	10%
Not FTB & 1 Adult, 1+ Child < 18	23%	10%	7%	5%	1%	2%
Not FTB & 2 Adults, No Child < 18	15%	17%	25%	27%	30%	30%
Not FTB & 3+ adults, No Child < 18	2%	6%	8%	8%	3%	3%
Not FTB & 2 Adults, 1+ Child < 18	25%	30%	33%	38%	46%	45%
Not FTB & Other	10%	10%	10%	10%	10%	10%

Table D.13: Education conditional prior probabilities.

<i>Income</i>	40k- 60k	60k- 80k	80k- 100k	100k- 120k	120k- 150k	150k+
Primary or below	20.0%	5.0%	1.0%	1.0%	0.0%	0.0%
Lower secondary	20.0%	10.0%	2.0%	1.0%	1.0%	0.0%
Higher secondary	25.0%	15.0%	4.0%	2.0%	1.0%	1.0%
Post leaving certificate	15.0%	25.0%	18.0%	10.0%	5.0%	3.0%
Third level non-degree	10.0%	28.0%	34.0%	28.0%	30.0%	30.0%
Third level degree or above	7.0%	14.0%	38.0%	55.0%	60.0%	63.0%
Other	3.0%	3.0%	3.0%	3.0%	3.0%	3.0%

Table D.14: Expenses-to-Household conditional prior probabilities.

<i>Household Composition</i>	<i>Household-to-Income</i>	<i>Variance</i>
1 Adult, No Child < 18	40.1%	5%
1 Adult, 1+ Child < 18	38.4%	10%
2 Adults, No Child < 18	48.9%	5%
3+ adults, No Child < 18	38.0%	10%
2 Adults, 1+ Child < 18	37.0%	5%
Other	45.1%	5%

Table D.15: Expenses-to-Income conditional prior probabilities.

<i>Income Group</i>	<i>Expenses-to-Income</i>	<i>Variance</i>
40k - 60k	54.1%	5.0%
60k - 80k	51.4%	5.0%
80k - 100k	47.7%	5.0%
100k - 120k	41.8%	6.0%
120k - 150k	38.0%	7.5%
150k+	31.3%	10.0%

D.3 Additional Default Settings

The default settings of a number of parameters are specified below. Any instances with feature values exceeding the maximum affordability or maximum MRTI are removed from the generated data. The generated continuous Loan Values are categorised based on Table D.16. House Value is calculated as Loan Value divided by Loan-to-Value. The generated House Values are then categorised based on Table D.17.

The Overall Default Rate for the generated data is specified at 2.75%. For this figure of 2.75%, Table D.18 specifies the distribution of the defaulters across the different risk groups.

The default Risk Scores for each Risk Group are specified in Table D.19.

Maximum Affordability = 11

Maximum MRTI = 0.8

Overall Default Rate = 2.75%

Noise = 0.25

Table D.16: Loan Value categories.

<i>Category</i>	<i>Start</i>	<i>End</i>
1	0.00	100k
2	100k	150k
3	150k	200k
4	200k	250k
5	250k	300k
6	300k	400k
7	400k	-

Table D.17: House Value categories.

<i>Category</i>	<i>Start</i>	<i>End</i>
1	0k	150k
2	150k	200k
3	200k	250k
4	250k	300k
5	300k	350k
6	350k	400k
7	400k	500k
8	500k	-

Table D.18: Distribution of the Overall Default Rate across the risk groups.

<i>Group</i>	<i>Default Rate</i>
1	70.0%
2	5.0%
3	2.5%
4	2.5%
5	2.5%
6	2.5%
7	2.5%
8	2.5%
9	2.5%
10	2.5%
11	1.5%
12	1.5%
13	1.0%
14	1.0%
15	0.0%

Table D.19: Risk level scores

<i>Risk Level</i>	<i>Risk Score</i>
1	0.953
2	7.885
3	11.939
4	14.816
5	17.047
6	18.871
7	20.412
8	21.748
9	22.925
10	23.026

References

- AL-GHAMDI, A. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention*, **34**, 729–741.
- ALAIZ-RODRÍGUEZ, R. & JAPKOWICZ, N. (2008). Assessing the impact of changing environments on classifier performance. In *Proceedings of the Canadian Society for Computational Studies of Intelligence 21st Conference on Advances in Artificial Intelligence*, 13–24, Springer-Verlag. 118, 201
- ALPAYDIN, E. (2004). *Introduction to machine learning*. The MIT Press, Cambridge, MA., USA, 2nd edn. 2
- ANDERSON, R. (2007). *The Credit Scoring Toolkit: Theory and Practice for Retail Credit Risk Management and Decision Automation*. Oxford University Press, USA. 51, 52, 53, 58, 71, 74, 77, 84, 88, 89, 96, 202

ANDERSSON, F., CHEN, Q. & GLENNON, D. (2011). Assessing the Impact of Changing Economic Conditions on the Design of Default Probability Models.

Presentation at Credit Scoring and Credit Control XII, Conference Proceedings Credit Research Centre, Business School, University of Edinburgh. 118

ANDREEVA, G. (2005). European generic scoring models using survival analysis.

Journal of the Operational research Society, **57**, 1180–1187. 112

ASUNCION, A. & NEWMAN, D. (2007). UCI Machine Learning Repository. University of California, Irvine, CA, School of Information and Computer Sciences. 113

ATIYA, A. (2001). Bankruptcy prediction for credit risk using neural networks: A survey and new results. *Neural Networks, IEEE Transactions on*, **12**, 929–935.

78

ATZMUELLER, M., BAUMEISTER, J., GOLLER, M. & PUPPE, F. (2006). A Data-generator for Evaluating Machine Learning Methods. *Journal Kunstliche Intelligenz*, **3**, 57–63. 119

BAESENS, B., VAN GESTEL, T., VIAENE, S., STEPANOVA, M., SUYKENS, J. & VANTHIENEN, J. (2003). Benchmarking state-of-the-art classification algorithms for credit scoring. *Journal of the Operational Research Society*, **54**, 627–635. 7, 103, 113, 139, 157

BAESENS, B., MUES, C., MARTEENS, D. & VANTHIENEN, J. (2009). 50 years of data mining and OR: upcoming trends and challenges. *Journal of the Operational Research Society*, **60**, 16–23. 5, 56, 70, 81

BALIN, B. (2008). Basel I, Basel II, and emerging markets: A nontechnical analysis. The Johns Hopkins University School of Advanced International Studies (SAIS), Washington, D.C.. Available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1477712, accessed on 28 January 2013. 59

BANASIK, J. & CROOK, J. (2009). Reject inference in survival analysis by augmentation. *Journal of the Operational Research Society*, **61**, 473–485. 87

BANASIK, J., CROOK, J. & THOMAS, L. (1996). Does scoring a subpopulation make a difference. *International Review of Retail, Distribution and Consumer Research*, **6**, 180–195. 90

BANASIK, J., CROOK, J. & THOMAS, L. (1999). Not if but when will borrowers default. *Journal of the Operational Research Society*, **50**, 1185–1190. 112

BANASIK, J., CROOK, J. & THOMAS, L. (2003). Sample selection bias in credit scoring models. *Journal of the Operational Research Society*, **54**, 822–832. 87

BASEL COMMITTEE ON BANKING SUPERVISION (1988). *International Convergence of Capital Measurement and Capital Standards (Updated to April 1998)*. Basel I, bank for International Settlements: Basel. Available at <http://www.bis.org/publ/bcbs04a.pdf>, last accessed accessed 30 January 2013. 58, 59

BASEL COMMITTEE ON BANKING SUPERVISION (2006). *International Convergence of Capital Measurement and Capital Standards - A Revised Framework*. Basel II, bank for International Settlements: Basel. Available at <http://www.bis.org/publ/bcbs128.pdf>, last accessed accessed 30 January 2013. 58, 63, 74, 95, 101

BASEL COMMITTEE ON BANKING SUPERVISION (2010). *International Convergence of Capital Measurement and Capital Standards (Updated to June 2011)*. Basel III, bank for International Settlements: Basel. Available at http://www.bis.org/publ/bcbs189_dec2010.pdf, last accessed accessed 30 January 2013. 58

BASEL COMMITTEE ON BANKING SUPERVISION (BCBS) (2001, Revised Edition 2005). *The Internal Ratings-Based Approach - Consultative Document*. Bank for International Settlements: Basel. Available at <http://www.bis.org/publ/bcbsca05.pdf>, last accessed accessed 30 January 2013. 64

BAWANEH, M., ALKOFFASH, M. & AL RABEA, A. (2008). Arabic Text Classification using K-NN and Naive Bayes. *Journal of Computer Science*, **4**, 600–605.

26

BELLOTTI, T. & CROOK, J. (2008). Credit scoring with macroeconomic variables using survival analysis. *Journal of the Operational Research Society*, **60**, 1699–1707. 29, 156

BELLOTTI, T. & CROOK, J. (2009). Support vector machines for credit scoring and discovery of significant features. *Expert Systems with Applications*, **36**, 3302–3308.

225

BENJAMIN, N., CATHCART, A. & RYAN, K. (2006). Low default portfolios: A proposal for conservative estimation of default probabilities. Tech. rep., Financial Services Authority: London, available at http://www.fsa.gov.uk/pubs/international/default_probabilities.pdf, last accessed 29 January 2013.

102

BERGKAMP, L. (2002). EU Data Protection Policy: The Privacy Fallacy: Adverse Effects of Europe's Data Protection Policy in an Information-Driven Economy.

Computer Law & Security Report, **18**, 31–47. 115

BEWICK, V., CHEEK, L. & BALL, J. (2005). Statistics review 14: Logistic regression. *Critical Care*, **9**, 112–118. 24

BIJAK, K. & THOMAS, L. (2012). Does segmentation always improve model performance in credit scoring? *Expert Systems with Applications*, **39**, 2433–2442. 4, 90, 111

BISHOP, C. (1994). Novelty detection and neural network validation. In *IEEE Proceedings: Vision, Image and Signal Processing*, vol. 141, 217–222, IEEE. 32

BISHOP, C. (1995). *Neural Networks for Pattern Recognition*. Oxford University Press, UK. 27, 35, 38

BISHOP, C. (2006). *Pattern recognition and machine learning*. Springer-Verlag, New York. xxi, 21, 22, 75

BRADLEY, A. (1997). The use of the area under the ROC curve in the evaluation of machine learning algorithms. *Pattern Recognition*, **30**, 1145–1159. 45, 225

BREEDEN, J., THOMAS, L. & McDONALD III, J. (2008). Stress testing retail loan portfolios with dual-time dynamics. *Journal of Risk Model Validation*, **2**, 43–62. 62

BREIMAN, L. (2001). Random forests. *Machine learning*, **45**, 5–32. 105

BREIMAN, L., FRIEDMAN, J., STONE, C. & OLSHEN, R. (1984). *Classification and regression trees*. Chapman & Hall/CRC, FL, USA. 90

BRITISH BANKERS ASSOCIATION (BBA) (2004). Introductory Paper on Low-Default Portfolios. available at <http://www.isda.org/speeches/pdf/ISDA-LIBA-BBA-LowDefaultPortfolioPaper080904-Introductory-Paper.pdf>, last accessed 29 January 2013. 101

BROWN, I. (2012). *Basel II compliant credit risk modelling: model development for imbalanced credit scoring data sets, loss given default (LGD) and exposure at default (EAD)*. Ph.D. thesis, University of Southampton. 105

BROWN, I. & MUES, C. (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets. *Expert Systems with Applications*, **39**, 3446–3453. 48, 105, 158

BUCHAN, I. (2011). Statsdirect (version 2.7.8)[computer software]. Cheshire, United Kingdom: Statsdirect. 178

BUREZ, J. & VAN DEN POEL, D. (2009). Handling class imbalance in customer churn prediction. *Expert Systems with Applications*, **36**, 4626–4636. 111, 157

BUTLER, W., STEWART, R. & MERRICK, G. (2009). A detailed radiobiological and dosimetric analysis of biochemical outcomes in a case-control study of permanent prostate brachytherapy patients. *Medical Physics*, **36**, 776. 177

CANBAS, S., CABUK, A. & KILIC, S. (2005). Prediction of commercial bank failure via multivariate statistical analysis of financial structures: The Turkish case. *European Journal of Operational Research*, **166**, 528–546. 80

CAO, Y. & HE, H. (2008). Learning from testing data: A new view of incremental semi-supervised learning. In *IEEE International Joint Conference on Neural Networks, IJCNN 2008 (IEEE World Congress on Computational Intelligence)*, 2872–2878, IEEE. 31

CARLIN, B. & LOUIS, T. (2008). *Bayesian methods for data analysis*. CRC Press, FL, USA, 3rd edn. 103

CARROLL, R. & RUPPERT, D. (1988). *Transformation and weighting in regression*. Chapman & Hall/CRC: London. 81

CASTELLANO, G. & FANELLI, A. (2000). Variable selection using neural-network models. *Neurocomputing*, **31**, 1–13. 81

CEBS (COMMITTEE OF EUROPEAN BANKING SUPERVISORS): VALIDATION GROUP (2005). Studies on the Validation of Internal Rating Systems (revised). Working Paper No. 14. available at http://www.bis.org/publ/bcbs_wp14.pdf, last accessed accessed 20 December 2012. 95

CHAN, K. & LOH, W. (2004). LOTUS: An algorithm for building accurate and comprehensible logistic regression trees. *Journal of Computational and Graphical Statistics*, **13**, 826–852. 90

CHANDLER, G. & COFFMAN, J. (1977). Using credit scoring to improve the quality of consumer receivables: Legal and statistical implications. In *Paper presented at the Financial Management Association meetings, Seattle, Washington*. 85

CHANDOLA, V., BANERJEE, A. & KUMAR, V. (2009). Anomaly detection: A survey. *ACM Computing Surveys (CSUR)*, **41**, 15. 32

CHANG, C.C. & LIN, C.J. (2001). LIBSVM: a library for support vector machines (version 2.9)[computer software]. Software available at <http://www.csie.ntu.edu.tw/~cjlin/libsvm>. 133

CHAPELLE, O., SCHÖLKOPF, B., ZIEN, A. *et al.* (2006). *Semi-supervised learning*. MIT press: Cambridge, MA. 31

CHAWLA, N., BOWYER, K., HALL, L. & KEGELMEYER, W. (2002). SMOTE: synthetic minority over-sampling technique. *Journal of Artificial Intelligence Research*, **16**, 321–357. 118, 239

CHAWLA, N.V., JAPKOWICZ, N. & KOTCZ, A. (2004). Special issue on learning from imbalanced data sets. *ACM SIGKDD Explorations Newsletter*, **6**, 1–6. 29, 30, 31, 107, 158

CHEN, G. & ASTEBRO, T. (2006). A maximum likelihood approach for reject inference in credit scoring. Tech. Rep. 07-05, Rotman School of Management Working Paper, available at http://papers.ssrn.com/sol3/papers.cfm?abstract_id=872541, last accessed 20 January 2013. 87

CHEN, G. & ÅSTEBRO, T. (2011). Bound and collapse Bayesian reject inference for credit scoring. *Journal of the Operational Research Society*, **63**, 1374–1387. 87

CHEN, S., HARDLE, W. & MORO, R. (2011). Modeling default risk with support vector machines. *Quantitative Finance*, **11**, 135–154. 7, 113

CHERKASSKY, V. & MULIER, F. (2007). *Learning from data: concepts, theory, and methods*. Wiley-IEEE Press: New York. 19

CHRISTENSEN, J., HANSEN, E. & LANDO, D. (2004). Confidence sets for continuous-time rating transition probabilities. *Journal of Banking & Finance*, **28**, 2575–2602. 101

CLEMENÇON, S., LUGOSI, G. & VAYATIS, N. (2005). Ranking and scoring using empirical risk minimization. *Learning Theory*, 783–800. 19

COOK, D. & HOLDER, L. (2001). A client-server interactive tool for integrated artificial intelligence curriculum. In *Proceedings of the Fourteenth International Florida Artificial Intelligence Research Society Conference AAAI Press*. 2

COVER, T. & HART, P. (1967). Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, **13**, 21–27. 37

CRONE, S. & FINLAY, S. (2012). Instance sampling in credit scoring: An empirical study of sample size and balancing. *International Journal of Forecasting*, **28**, 224–238. 71, 72, 111, 156

CROOK, J. & BANASIK, J. (2004). Does reject inference really improve the performance of application scoring models? *Journal of Banking & Finance*, **28**, 857–874. 87

CROOK, J., EDELMAN, D. & THOMAS, L. (2007). Recent developments in consumer credit risk assessment. *European Journal of Operational Research*, **183**, 1447–1465. 52, 94, 95

CSO (2010). Statistical Yearbook of Ireland 2010 Edition. available at <http://www.cso.ie/en/releasesandpublications/othercsopublications/>

[statisticalyearbookofireland2010edition/](http://www.statisticsireland.ie/2010edition/), last accessed 29 January 2013.

200, 202, 203, 204, 207

CUNNINGHAM, P., CORD, M. & DELANY, S. (2008). Supervised learning. *Machine Learning Techniques for Multimedia*, 21–49. 18

DEMŠAR, J. (2006). Statistical comparisons of classifiers over multiple data sets. *The Journal of Machine Learning Research*, 7, 1–30. 48, 49, 128

DIAMANTOPOULOS, A. & SIGUAW, J. (2006). Formative versus reflective indicators in organizational measure development: A comparison and empirical illustration. *British Journal of Management*, 17, 263–282. 78

DIAMOND, C. & SIMON, C. (1990). Industrial specialization and the returns to labor. *Journal of Labor Economics*, 175–201. 80

DIETTERICH, T. (1998). Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10, 1895–1923. 40

DIONNE, G., ARTÍS, M. & GUILLÉN, M. (1996). Count data models for a credit scoring system. *Journal of Empirical Finance*, 3, 303–325. 126

DOFE (2008). Department of the Environment, Community and Local Government: Latest House Prices, Loans and Profile of Borrowers Statistics. available at <http://www.environ.ie/en/Publications/StatisticsandRegularPublications/HousingStatistics/>, last accessed 29 January 2013. 167, 200, 202, 203, 204, 205, 207

DRUMMOND, C. & HOLTE, R. (2000). Explicitly representing expected cost: An alternative to ROC representation. In *Proceedings of the 6th ACM SIGKDD international conference on knowledge discovery and data mining*, 198–207, ACM.

47

DRUMMOND, C. & HOLTE, R. (2004). C4.5, class imbalance, and cost sensitivity: Why under-sampling beats over-sampling. In *Proceedings of the 20th International Conference on Machine Learning (ICML'03) Workshop on Learning from Imbalanced Data Sets II*. 158

DRUMMOND, C. & HOLTE, R. (2005a). Learning to live with false alarms. In *Workshop on Data Mining Methods for Anomaly Detection held in conjunction with the 11th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 21–24. 114

DRUMMOND, C. & HOLTE, R. (2005b). Severe class imbalance: Why better algorithms aren't the answer. *Machine Learning: ECML 2005*, 539–546. 29

DRUMMOND, C. & HOLTE, R.C. (2006). Cost curves: An improved method for visualizing classifier performance. *Machine Learning*, **65**, 95–130. 47

DRUMMOND, C. & JAPKOWICZ, N. (2010). Warning: Statistical benchmarking is addictive. Kicking the habit in machine learning. *Journal of Experimental & Theoretical Artificial Intelligence*, **22**, 67–80. 114

DUDA, R.O. & HART, P.E. (1973). *Pattern classification and scene analysis (1st edition)*. Wiley: New York. 26

DUIN, R. (1976). On the choice of smoothing parameters for parzen estimators of probability density functions. *IEEE Transactions on Computers*, **100**, 1175–1179.

36

DUIN, R., JUSZCZAK, P., PACLIK, P., PEKALSKA, E., DE RIDDER, D., TAX, D. & VERZAKOV, S. (2008). *PRTOOLS (V4.1.4): A Matlab toolbox for pattern recognition*. Delft University of Technology, NL. 133

DWYER, D. (2007). The distribution of defaults and Bayesian model validation. *The Journal of Risk Model Validation*, **1**, 23–53. 102

EUROSTAT (2008). NACE Rev. 2, Statistical classification of economic activities in the European Community. Available at http://epp.eurostat.ec.europa.eu/portal/page/portal/nace_rev2/correspondence_tables, last accessed 29 January 2013. 203

FAWCETT, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, **27**, 861–874. 45

FEELDERS, A. (2000). Credit scoring and reject inference with mixture models. *International Journal of Intelligent Systems in Accounting, Finance & Management*, **9**, 1–8. 87

FINLAY, S. (2011). Multiple classifier architectures and their application to credit risk assessment. *European Journal of Operational Research*, **210**, 368–378. 57

FISCHER, B. & ZIGMOND, M. (2010). The essential nature of sharing in science. *Science and Engineering Ethics*, **16**, 783–799. 114

FISHER, R. (1936). The use of multiple measurements in taxonomic problems. *Annals of Human Genetics*, **7**, 179–188. 21

FITCH RATINGS (2007). AIB Mortgage Bank Mortgage Covered Securities. Available at <http://www.fitchratings.com>, last accessed 29 January 2013. 202, 207

FLACH, P., HERNÁNDEZ-ORALLO, J. & FERRI, C. (2011). A coherent interpretation of AUC as a measure of aggregated classification performance. In *Proceedings of the 28th International Conference on Machine Learning (ICML 2011)*, 657–664.

47

FLOREZ-LOPEZ, R. (2009). Effects of missing data in credit risk scoring. A comparative analysis of methods to achieve robustness in the absence of sufficient data. *Journal of the Operational Research Society*, **61**, 486–501. 70, 123

FOGARTY, D. (2006). Multiple imputation as a missing data approach to reject inference on consumer credit scoring. Tech. rep., University of Phoenix, available at <http://interstat.statjournals.net/YEAR/2006/articles/0609001.pdf> last accessed 28 January 2013. 87

FORREST, A. (2005). Likelihood approaches to low default portfolios. In *Credit Scoring and Credit Control IX, Conference Proceedings, Credit Research Centre, Business School, University of Edinburgh*, CRC. 102

FRIEDMAN, J. (2001). Greedy function approximation: a gradient boosting machine.(english summary). *Annals of Statistics*, **29**, 1189–1232. 105

FRIEDMAN, J. (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**, 367–378. 105

FRIEDMAN, J., HASTIE, T. & TIBSHIRANI, R. (2001). *The elements of statistical learning*. Springer Series in Statistics, NY, USA, 1st edn. 19, 36

FRIEDMAN, M. (1937). The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the American Statistical Association*, **32**, 675–701. 49, 128

GARCIA, S. & HERRERA, F. (2008). An extension on statistical comparisons of classifiers over multiple data sets for all pairwise comparisons. *Journal of Machine Learning Research*, **9**, 2677–2694. 49

GARCÍA, S., FERNÁNDEZ, A., LUENGO, J. & HERRERA, F. (2010). Advanced nonparametric tests for multiple comparisons in the design of experiments in computational intelligence and data mining: Experimental analysis of power. *Information Sciences*, **180**, 2044–2064. 49

GHASEMI, A., MANZURI, M.T., RABIEE, H.R., ROHBAN, M.H. & HAGHIRI, S. (2011). Active one-class learning by kernel density estimation. In *Machine Learning for Signal Processing (MLSP), 2011 IEEE International Workshop on*, 1–6, IEEE. 239

GREENSPAN, A. (2002). Remarks by U.S. Federal Reserve Chairman Alan Greenspan. Speech at the American Bankers Association, Phoenix, Arizona. Available at <http://www.federalreserve.gov/boarddocs/speeches/2002/20021007/default.htm>, last accessed 29 January 2013. 56

GUYON, I. & ELISSEEFF, A. (2003). An introduction to variable and feature selection. *The Journal of Machine Learning Research*, **3**, 1157–1182. 75

HAENLEIN, M., KAPLAN, A. & BEESER, A. (2007). A model to determine customer lifetime value in a retail banking context. *European Management Journal*, **25**, 221–234. 56

HAND, D. (2001). Modelling consumer credit risk. *IMA Journal of Management Mathematics*, **12**, 139–155. 53, 55, 56, 177

HAND, D. (2005). Good practice in retail credit scorecard assessment. *Journal of the Operational Research Society*, **56**, 1109–1117. 45

HAND, D. (2006a). Classifier technology and the illusion of progress. *Statistical Science*, **21**, 1–14. 158

HAND, D. (2006b). Rejoinder: Classifier Technology and the Illusion of Progress. *Statistical Science*, **21**, 30–34. 223

HAND, D. (2009). Measuring classifier performance: a coherent alternative to the area under the ROC curve. *Machine Learning*, **77**, 103–123. 46, 126, 128

HAND, D. (2012). Assessing the performance of classification methods. *International Statistical Review*, **80**, 400–414. 91

HAND, D. & ADAMS, N. (2000). Defining attributes for scorecard construction in credit scoring. *Journal of Applied Statistics*, **27**, 527–540. 117

HAND, D. & HENLEY, W. (1993). Can reject inference ever work? *IMA Journal of Management Mathematics*, **5**, 45–55. 85

HAND, D. & HENLEY, W. (1997). Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society: Series A (Statistics in Society)*, **160**, 523–541. 3, 56, 57, 108

HAND, D. & YU, K. (2001). Idiot's Bayes: Not So Stupid after All? *International Statistical Review*, **69**, 385–398. 26

HAND, D. & ZHOU, F. (2009). Evaluating models for classifying customers in retail banking collections. *Journal of the Operational Research Society*, **61**, 1540–1547. 23, 91, 124, 180

HAND, D., MANNILA, H. & SMYTH, P. (2001). *Principles of data mining*. MIT Press, MA, USA. 88

HAND, D., SOHN, S. & KIM, Y. (2005). Optimal bipartite scorecards. *Expert Systems with Applications*, **29**, 684–690. 81, 82

HANLEY, J. & MCNEIL, B. (1982). The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, **143**, 29–36. 45

HANSON, S. & SCHUERMANN, T. (2006). Confidence intervals for probabilities of default. *Journal of Banking & Finance*, **30**, 2281–2301. 101

HAWKINS, D. *et al.* (2004). The problem of overfitting. *Journal of chemical information and computer sciences*, **44**, 1–12. 30

HEMPSTALK, K. (2009). *Continuous Typist Verification using Machine Learning*. Ph.D. thesis, The University of Waikato, New Zealand. 3, 32, 34

- HENLEY, W. & HAND, D. (1996). A k-nearest-neighbour classifier for assessing consumer credit risk. *The Statistician*, **45**, 77–95. 28
- HINTON, G. (1989). Connectionist learning procedures. *Artificial intelligence*, **40**, 185–234. 38
- HOADLEY, B. (2001). [Statistical Modeling: The Two Cultures]: Comment. *Statistical Science*, **16**, 220–224. 73
- HOCKING, R. (1976). A biometrics invited paper: The analysis and selection of variables in linear regression. *Biometrics*, **32**, 1–49. 79
- HODGES, J. & LEHMANN, E. (1962). Rank methods for combination of independent experiments in analysis of variance. *The Annals of Mathematical Statistics*, **33**, 482–497. 49
- HOFFMANN, F., BAESENS, B., MUES, C., VAN GESTEL, T. & VANTHIENEN, J. (2007). Inferring descriptive and approximate fuzzy rules for credit scoring using evolutionary algorithms. *European Journal of Operational Research*, **177**, 540–555. 104, 117
- HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics*, **6**, 65–70. 128
- HOLMAN, J. (2010). Flawed solution: The difficulties of mandating a leverage ratio in the United States. *Southern California Law Review*, **84**, 713–751. 59
- HOSMER, D. & LEMESHOW, S. (2000). *Applied logistic regression*. Wiley-Interscience: New York, 2nd edn. 23, 79

HUANG, Z., CHEN, H., HSU, C., CHEN, W. & WU, S. (2004). Credit rating analysis with support vector machines and neural networks: a market comparative study. *Decision Support Systems*, **37**, 543–558. 81

JAPKOWICZ, N. (1999). *Concept-learning in the absence of counter-examples: An autoassociation-based approach to classification*. Ph.D. thesis, Rutgers, The State University of New Jersey. 32, 38

JAPKOWICZ, N. (2000). The class imbalance problem: Significance and strategies. In *Proceedings of the 2000 International Conference on Artificial Intelligence (ICAI 2000)*, vol. 1, 111–117. 29

JAPKOWICZ, N. & SHAH, M. (2011). *Evaluating Learning Algorithms: A Classification Perspective*. Cambridge University Press, 1st edn. 39, 47, 75, 116, 230

JOANES, D. (1993). Reject inference applied to logistic regression for credit scoring. *IMA Journal of Management Mathematics*, **5**, 35–43. 87

JOHN, G. & LANGLEY, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *Proceedings of the 11th conference on uncertainty in artificial intelligence*, 338–345, Morgan Kaufmann Publishers Inc. 26

JUSZCZAK, P., ADAMS, N., HAND, D., WHITROW, C. & WESTON, D. (2008). Off-the-peg and bespoke classifiers for fraud detection. *Computational Statistics and Data Analysis*, **52**, 4521–4532. 3, 32, 106, 238

KASS, G. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics*, **29**, 119–127. 90

KELLY, M., HAND, D. & ADAMS, N. (1999). The impact of changing populations on classifier performance. In *Proceedings of the 5th ACM SIGKDD international conference on knowledge discovery and data mining*, 367–371, ACM. 117, 224

KELLY, R., MCCARTHY, Y. & MCQUINN, K. (2012). Impairment and Negative Equity in the Irish Mortgage Market. *Journal of Housing Economics*, **21**, 256–268. 200, 213

KENNEDY, K., MAC NAMEE, B. & DELANY, S. (2010). Learning without default: A study of one-class classification and the low-default portfolio problem. In *Artificial Intelligence and Cognitive Science: 20th Irish Conference, AICS 2009, Dublin, Ireland, 2009, Revised Selected Papers*, vol. 6206, 174–187, Springer. 13

KENNEDY, K., DELANY, S. & MAC NAMEE, B. (2011). A framework for generating data to simulate application scoring. In *Credit Scoring and Credit Control XII, Conference Proceedings, Credit Research Centre, Business School, University of Edinburgh*, CRC. 10, 13, 14

KENNEDY, K., DELANY, S. & MAC NAMEE, B. (2012a). An artificial data generation framework for retail credit scoring problems: Technical addendum. 207, 213, 225

KENNEDY, K., MAC NAMEE, B. & DELANY, S. (2012b). Using semi-supervised classifiers for credit scoring. *Journal of the Operational Research Society*, **64**, 513–529. 9, 13, 14

KENNEDY, K., MAC NAMEE, B., DELANY, S., O'SULLIVAN, M. & WATSON, N.

(2012c). A window of opportunity: Assessing behavioural scoring. *Expert Systems with Applications*, **40**, 1372–1380. 10, 13, 14

KEOGH, E. (2007). Why the lack of reproducibility is crippling research in data mining and what you can do about it. In *Proceedings of the 8th international workshop on Multimedia data mining: (associated with the ACM SIGKDD 2007)*, 2, ACM. 127

KHOSHGOFTAAR, T., SEIFFERT, C., VAN HULSE, J., NAPOLITANO, A. & FOLECO, A. (2007). Learning with limited minority class data. In *6th International Conference on Machine Learning and Applications, 2007 (ICMLA 2007)*, 348–353, IEEE. 239

KIEFER, N. (2009). Default estimation for low-default portfolios. *Journal of Empirical Finance*, **16**, 164–173. 102

KIM, J. (2009). Estimating classification error rate: Repeated cross-validation, repeated hold-out and bootstrap. *Computational Statistics & Data Analysis*, **53**, 3735–3745. 47

KLEINBAUM, D. & KLEIN, M. (2010). Maximum likelihood techniques: An overview. *Logistic Regression*, 103–127. 24

KOLB, R. & OVERDAHL, J. (2009). *Financial Derivatives: Pricing and Risk Management*. Wiley, USA, 5th edn. 59

KOLCZ, A., CHOWDHURY, A. & ALSPECTOR, J. (2003). Data duplication: An imbalance problem. In *Proceedings of the ICML 2003 workshop on learning from imbalanced datasets (II), Washington, DC, USA*. 142

KRIVKO, M. (2010). A hybrid model for plastic card fraud detection systems. *Expert Systems with Applications*, **37**, 6070–6076. 106

KRUSKAL, W. & WALLIS, W. (1952). Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association*, 583–621. 49

LANGLEY, P. & SIMON, H. (1995). Applications of machine learning and rule induction. *Communications of the ACM*, **38**, 54–64. 28

LE CESSIE, S. & VAN HOUWELINGEN, J. (1992). Ridge estimators in logistic regression. *Applied Statistics*, 191–201. 225

LE CUN, Y., DENKER, J.S., SOLLA, S.A., HOWARD, R.E. & JACKEL, L.D. (1990). Optimal brain damage. *Advances in neural information processing systems*, **2**, 598–605. 132

LEE, H. & CHO, S. (2007). Focusing on non-respondents: Response modeling with novelty detectors. *Expert Systems with Applications*, **33**, 522–530. 56, 157

LEE, S. (2005). Application of logistic regression model and its validation for landslide susceptibility mapping using gis and remote sensing data. *International Journal of Remote Sensing*, **26**, 1477–1491. 25

LEE, T. & CHEN, I. (2005). A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, **28**, 743–752. 4

LESSMANN, S., BAESENS, B., MUES, C. & PIETSCH, S. (2008). Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Transactions on Software Engineering*, **34**, 485–496. 43

LEUNG KAN HING, K. (2008). *An Investigation of Artificial Immune Systems and Variable Selection Techniques for Credit Scoring*. Ph.D. thesis, Royal Melbourne Institute of Technology University, Australia. 78, 80, 202

LEWIS, E. (1992). *An introduction to credit scoring*. Athena Press: San Rafael, CA.

71

LIN, S., ANSELL, J. & ANDREEVA, G. (2011). Predicting default of a small business using different definitions of financial distress. *Journal of the Operational Research Society*, **63**, 539–548. 82, 85

LINDSAY, R., JACKSON, T. & COOKE, L. (2010). Mobile access to information systems in law enforcement: An evaluation of its implications for data quality. *Electronic Journal Information Systems Evaluation Volume*, **13**, 143–152. 70

LIU, Y. & SCHUMANN, M. (2005). Data mining feature selection for credit scoring models. *Journal of the Operational Research Society*, **56**, 1099–1108. 133

LOOG, M. & DUIN, R. (2002). Non-iterative heteroscedastic linear dimension reduction for two-class data. In *Proceedings of the Joint IAPR International*

Workshop on Structural, Syntactic, and Statistical Pattern Recognition, 508–517,

Springer-Verlag. 22

LOUGHREY, J. & CUNNINGHAM, P. (2005). Overfitting in wrapper-based feature subset selection: The harder you try the worse it gets. In *Research and Development in Intelligent Systems XXI: Proceedings of AI-2004, the 24th Sgai International Conference on Innovative Techniques and Applications of Artificial Intelligence*, 33–43, Springer-Verlag New York Incorporated. 76

LYDON, R. & McCARTHY, Y. (2011). What lies beneath? Understanding recent trends in Irish mortgage arrears. Research Technical Papers, Central Bank of Ireland, available at <http://www.centralbank.ie/publications/Documents/14RT11.pdf>, last accessed 29 January 2013. 200, 213

MALIK, M. & THOMAS, L. (2009). Modelling credit risk of portfolio of consumer loans. *Journal of the Operational Research Society*, **61**, 411–420. 64, 108

MALIK, M. & THOMAS, L. (2012). Transition matrix models of consumer credit ratings. *International Journal of Forecasting*, **28**, 261–272. 109

MALIN, S. & SCHLAPP, D. (1980). Geomagnetic lunar analysis by least-squares. *Geophysical Journal of the Royal Astronomical Society*, **60**, 409–418. 115

MARQUÉS, A., GARCÍA, V. & SÁNCHEZ, J. (2012). On the suitability of resampling techniques for the class imbalance problem in credit scoring. *Journal of the Operational Research Society*, **10.1057/jors.2012.120**. 157, 158

MARTENS, D., BAESENS, B., GESTEL, T.V. & VANTHIENEN, J. (2007). Comprehensible credit scoring models using rule extraction from support vector machines.

European Journal of Operational Research, **183**, 1466–1476. 117

MARTENS, D., VAN GESTEL, T., DE BACKER, M., HAESEN, R., VANTHIENEN, J. & BAESENS, B. (2010). Credit rating prediction using ant colony optimization.

Journal of the Operational Research Society, **61**, 561–573. 72

MARTENS, D., BAESENS, B. & FAWCETT, T. (2011). Editorial survey: Swarm intelligence for data mining. *Machine Learning*, **83**, 1–42. 113

MAYS, E. (2004). *Credit scoring for risk managers: The handbook for lenders*. Thomson/South-Western, OH, USA. 52, 56, 76, 84, 87, 90, 110, 165, 173

MCCARTHY, J., MINSKY, M., ROCHESTER, N. & SHANNON, C. (1955). A proposal for the dartmouth summer research project on artificial intelligence. Tech. rep., Dartmouth College, <http://www.aaai.org/ojs/index.php/aimagazine/article/viewArticle/1904> last accessed 29/01/2013. 2

MCCARTHY, Y. & MCQUINN, K. (2010). How are Irish households coping with their mortgage repayments? Information from the SILC Survey. Research Technical Papers, Central Bank & Financial Services Authority of Ireland (CBFSAI), available at <http://www.centralbank.ie/publications/documents/2RT10.pdf>, last accessed 29 January 2013. 210, 213

MCDONALD, R., STURGESSION, M., SMITH, K., HAWKINS, M. & HUANG, E. (2012). Non-linearity of scorecard log-odds. *International Journal of Forecasting*, **28**, 239–247. 108

MCNAB, H. & WYNN, A. (2000). *Principles and practice of consumer credit risk management*. Chartered Institute of Bankers Publishing, Canterbury. xiv, 73, 108, 109, 110

MEEHL, P. (1955). Clinical versus statistical prediction. *Journal of Consulting Psychology*, **19**, 155. 104

MEESTER, S. (2000). Reject inference for credit scoring model development using extrapolation. Tech. rep., Mimeo, CIT Group, NJ, USA. 86

MELLI, G. (2007). Dataset Generator [Computer Software], available at <http://www.datasetgenerator.com>. 119

MERCER, J. (1909). Functions of positive and negative type, and their connection with the theory of integral equations. *Philosophical Transactions of the Royal Society of London. Series A, Containing Papers of a Mathematical or Physical Character*, **209**, 415–446. 27

MICHIE, D., SPIEGELHALTER, D., TAYLOR, C. & CAMPBELL, J. (1994). *Machine learning, neural and statistical classification*. Ellis Horwood, UK. 103

MILLER, A.J. (1984). Selection of subsets of regression variables. *Journal of the Royal Statistical Society. Series A (General)*, 389–425. 181

MIN, J. & LEE, Y. (2005). Bankruptcy prediction using support vector machine with optimal choice of kernel function parameters. *Expert Systems with Applications*, **28**, 603–614. 80

MIRA, J. (2008). Symbols versus connections: 50 years of artificial intelligence.

Neurocomputing, **71**, 671–680. 2

MITCHELL, T. (1997). *Machine learning*. 2

MOODY'S (2010a). Celtic Residential Irish Mortgage Securitisation No. 16 Limited.

Available at <http://www.moodys.com>, last accessed 24th March 2011. 202, 205

MOODY'S (2010b). What Drives Irish Mortgage Borrowers to Default. Available at

<http://www.alacrastore.com/>. 200, 207, 210, 213, 215, 216

MORRISON, J. (2004). Variable selection in model development. Tech. rep., Tran-

sUnion, Atlanta. 78

MOSLEY, L. & SINGER, D. (2009). The global financial crisis: Lessons and opportu-

nities for international political economy. *International Interactions*, **35**, 420–429.

57

MOYA, M., KOCH, M. & HOSTETLER, L. (1993). One-class classifier networks

for target recognition applications. Tech. rep., SAND-93-0084C, Sandia National

Labs., Albuquerque, NM (United States). 31, 32

MOŽINA, M., ŽABKAR, J. & BRATKO, I. (2007). Argument based machine learn-

ing. *Artificial Intelligence*, **171**, 922–937. 126

MULLER, K., MIKA, S., RATSCH, G., TSUDA, K. & SCHOLKOPF, B. (2001). An

introduction to kernel-based learning algorithms. *IEEE Transactions on Neural*

Networks, **12**, 181–201. 19

MWANGI, I. & SICHEI, M. (2011). Determinants of access to credit by individuals in Kenya: A comparative analysis of the kenya national finaccess surveys of 2006 and 2009. *European Journal of Business and Management*, **3**, 206–226. 77

MYERS, G. (1999). A dataset generator for whole genome shotgun sequencing. In *Proceedings of the 7th International Conference on Intelligent Systems for Molecular Biology*, 202–210, AAAI Press. 118

MYERS, R. (1990). *Classical and modern regression with applications*. Duxbury Press Belmont, CA, 2nd edn. 79

NANNI, L. & LUMINI, A. (2009). An experimental comparison of ensemble of classifiers for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, **36**, 3028–3033. 126

OeNB/FMA (2004). Guidelines on Credit Risk Management –Rating Models and Validation. Available at http://www.oenb.at/en/img/rating_models_tcm16-22933.pdf, last accessed 20 May 2013. 104

ORTH, W. (2011). Default probability estimation in small samples-with an application to sovereign bonds. Tech. rep., University of Cologne, discussion Papers in Statistics and Econometrics 05/2011. 102, 103

OVERSTREET, G., BRADLEY, E. & KEMP, R. (1992). The flat-maximum effect and generic linear scoring models: A test. *IMA Journal of Management Mathematics*, **4**, 97. 158

OZKAYA, S. & SIYABI, S. (2008). Detection of fracture corridors from dynamic data by factor analysis. In *Society of Petroleum Engineers, Saudi Arabia, Technical Symposium Section*. 80

PARK, S., MURPHY, S., WILKENS, L., YAMAMOTO, J., SHARMA, S., HANKIN, J., HENDERSON, B. & KOLONEL, L. (2005). Dietary patterns using the food guide pyramid groups are associated with sociodemographic and lifestyle factors: the multiethnic cohort study. *The Journal of Nutrition*, **135**, 843–849. 80

PARKER, M., MOLESHE, V., DE LA HARPE, R. & WILLS, G. (2006). An evaluation of information quality frameworks for the world wide web. In *Proceedings of the 8th Annual Conference of WWW Applications*. 70

PARZEN, E. (1962). On estimation of a probability density function and mode. *The Annals of Mathematical Statistics*, **33**, 1065–1076. 35

PAVLIDIS, N., TASOULIS, D., ADAMS, N. & HAND, D. (2012). Adaptive consumer credit classification. *Journal of the Operational Research Society*, **63**, 1645–1654. 224

PEARCE, J. & FERRIER, S. (2000). An evaluation of alternative algorithms for fitting species distribution models using logistic regression. *Ecological Modelling*, **128**, 127–147. 79

PHUA, C., LEE, V., SMITH, K. & GAYLER, R. (2010). A comprehensive survey of data mining-based fraud detection research. *CoRR*, **abs/1009.6119**. 56

PIETRUSZKIEWICZ, W. (2008). Dynamical systems and nonlinear Kalman filtering applied in classification. In *Proceedings of the 7th IEEE International Conference on Cybernetic Intelligent Systems*, 263–268, IEEE. 126

PIRAMUTHU, S. (1999). Financial credit-risk evaluation with neural and neurofuzzy systems. *European Journal of Operational Research*, **112**, 310–321. 132

PLUTO, K. & TASCHE, D. (2006). Estimating probabilities of default for low default portfolios. *The Basel II Risk Parameters*, 79–103. 101, 102

PLUTO, K. & TASCHE, D. (2011). Estimating probabilities of default for low-default portfolios. *The Basel II Risk Parameters*, 75–101. 101

PROVOST, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI 2000 Workshop on Imbalanced Data Sets*. 30, 149

RATANAMAHATANA, C. & GUNOPULOS, D. (2003). Feature selection for the naïve Bayesian classifier using decision trees. *Applied Artificial Intelligence*, **17**, 475–487. 81

RIPLEY, B. (1994). Neural networks and related methods for classification. *Journal of the Royal Statistical Society. Series B (Methodological)*, **56**, 409–456. 117

RITTER, G. & GALLEGOS, M.T. (1997). Outliers in statistical pattern recognition and an application to automatic chromosome classification. *Pattern Recognition Letters*, **18**, 525–540. 32

ROKACH, L. (2010). *Pattern classification using ensemble methods*. World Scientific, Singapore. 17

ROMASCO, A. (1983). *The politics of recovery: Roosevelt's New Deal*. Oxford University Press, NY, USA. 54

SABZEVARI, H., SOLEYMANI, M. & NOORBAKSH, E. (2007). A comparison between statistical and data mining methods for credit scoring in case of limited available data. 126

SAITTA, L. & NERI, F. (1998). Learning in the real world. *Machine Learning*, **30**, 133–163. 114

SALZBERG, S.L. (1997). On comparing classifiers: Pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, **1**, 317–328. 113

SARLIJA, N., BENSIC, M. & ZEKIC-SUSAC, M. (2009). Comparison procedure of predicting the time to default in behavioural scoring. *Expert Systems with Applications*, **36**, 8778–8788. 108

SARMIENTO, T., HONG, S. & MAY, G. (2005). Fault detection in reactive ion etching systems using one-class support vector machines. In *2005 IEEE/SEMI Advanced Semiconductor Manufacturing Conference and Workshop*, 139–142. 3, 32

SCOTT, P. & WILKINS, E. (1999). Evaluating data mining procedures: Techniques for generating artificial data sets. *Information and software technology*, **41**, 579–587. 116, 118, 201

SEBASTIANI, P. & RAMONI, M. (2000). Bayesian inference with missing data using bound and collapse. *Journal of Computational and Graphical Statistics*, **9**, 779–800. 87

SEIFFERT, C., KHOSHGOFTAAR, T. & VAN HULSE, J. (2009). Improving software-quality predictions with data sampling and boosting. *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, **39**, 1283–1294. 29

SELIYA, N., KHOSHGOFTAAR, T.M. & VAN HULSE, J. (2009). A study on the relationships of classifier performance metrics. In *Tools with Artificial Intelligence, 2009. ICTAI'09. 21st International Conference on*, 59–66, IEEE. 46

SHawe-Taylor, J., Bartlett, P., Williamson, R. & Anthony, M. (1998). Structural risk minimization over data-dependent hierarchies. *IEEE Transactions on Information Theory*, **44**, 1926–1940. 19

SHERLOCK, G. et al. (2000). Analysis of large-scale gene expression data. *Current opinion in immunology*, **12**, 201–205. 90

SHESKIN, D. (1997). *Handbook of parametric and nonparametric statistical procedures*. CRC Press, FL, USA. 177

SHIN, K. & LEE, Y. (2002). A genetic algorithm application in bankruptcy prediction modeling. *Expert Systems with Applications*, **23**, 321–328. 81

SHIN, K., LEE, T. & KIM, H. (2005). An application of support vector machines in bankruptcy prediction model. *Expert Systems with Applications*, **28**, 127–135.

81

SHLAY, A. (2006). Low-income homeownership: American dream or delusion? *Urban Studies*, **43**, 511–531. 54

SIDDIQI, N. (2005). *Credit risk scorecards: Developing and implementing intelligent credit scoring*. John Wiley & Sons: Hoboken, NJ, USA. 64, 65, 71, 73, 74, 76, 82, 84, 85, 89, 92, 93

SIEGEL, S. (1956). *Nonparametric statistics for the behavioral sciences*. McGraw-Hill, Tokyo. 95

SOARES, C. (2003). Is the UCI repository useful for data mining? *Progress in Artificial Intelligence*, 209–223. 113, 114

SOBEHART, J., KEENAN, S. & STEIN, R. (2000). Benchmarking quantitative default risk models: a validation methodology. *Moody's Investors Service*. 95

SOHN, S. & SHIN, H. (2006). Reject inference in credit operations based on survival analysis. *Expert Systems with Applications*, **31**, 26–29. 87

SPRINTHAL, R. & FISK, S. (1990). *Basic statistical analysis*. Prentice Hall, NJ, USA. 95

SRIKANT, R. (1994). Quest synthetic data generation code. San Jose: IBM Almaden Research Center. 118, 201

STEFANESCU, C., TUNARU, R. & TURNBULL, S. (2009). The credit rating process and estimation of transition probabilities: A Bayesian Approach. *Journal of Empirical Finance*, **16**, 216–234. 102, 104

STEIN, R. (2002). Benchmarking default prediction models: Pitfalls and remedies in model validation. Tech. Rep. 20305, Moody's KMV, New

York, available at https://riskcalc.moodysrms.com/us/research/crm/validation_tech_report_020305.pdf, last accessed 29 January 2013. 94, 95

SUYKENS, J. & VANDEWALLE, J. (1999). Least squares support vector machines. *Neural processing letters*, **9**, 293–300. 105

TANG, T. & CHI, L. (2005). Predicting multilateral trade credit risks: Comparisons of logit and fuzzy logic models using ROC curve analysis. *Expert Systems with Applications*, **28**, 547–556. 104

TASCHE, D. (2012). Bayesian estimation of probabilities of default for low default portfolios. Tech. rep., Financial Services Authority, UK, available at SSRN: <http://ssrn.com/abstract=2048818>, Last accessed 29 January 2013. 102

TAX, D. (2001). *One-class classification*. Ph.D. thesis, Delft University of Technology, NL. 32, 33, 34, 35, 37

TAX, D. (2009). DDtools, the Data Description Toolbox for Matlab (version 2.7.8)[computer software]. 34, 133

TAX, D. & DUIN, R. (1999). Support vector domain description. *Pattern Recognition Letters*, **20**, 1191–1199. 33, 37

TAX, D. & DUIN, R. (2000). Data description in subspaces. In *Proceedings of 15th IEEE International Conference on Pattern Recognition*, vol. 2, 672–675, IEEE. 37

THOMAS, J. (2001). A methodology for linking customer acquisition to customer retention. *Journal of Marketing Research*, **32**, 262–268. 56

THOMAS, L. (2009a). *Consumer credit models: Pricing, profit, and portfolios*. Oxford University Press, USA. 52, 56, 62, 65, 75, 76, 79, 82, 83, 86, 88, 93, 111

THOMAS, L. (2009b). Operations research in consumer finance: Challenges for operational research. *Journal of Operational Research Society*, **61**, 41–52. 5, 52, 54, 103, 124, 127, 158, 238

THOMAS, L., BANASIK, J. & CROOK, J. (2001a). Recalibrating scorecards. *Journal of the Operational Research Society*, **52**, 981–988. 92

THOMAS, L., HO, J. & SCHERER, W. (2001b). Time will tell: Behavioural scoring and the dynamics of consumer credit assessment. *IMA Journal of Management Mathematics*, **12**, 89–103. 109, 111, 176, 187, 194

THOMAS, L., OLIVER, R. & HAND, D. (2005). A survey of the issues in consumer credit modelling research. *Journal of the Operational Research Society*, **56**, 1006–1015. 52

THOMAS, L.C., EDELMAN, D.B. & CROOK, J.N. (2002). *Credit scoring and its applications*. Society for Industrial and Applied Mathematics, Philadelphia, USA. 24, 53, 85, 86, 126, 248

TINSLEY, H. & TINSLEY, D. (1987). Uses of factor analysis in counseling psychology research. *Journal of counseling psychology*, **34**, 414. 80

TSAI, C. (2009). Feature selection in bankruptcy prediction. *Knowledge-Based Systems*, **22**, 120–127. 111, 126

TSAI, C. & WU, J. (2008). Using neural network ensembles for bankruptcy prediction and credit scoring. *Expert Systems with Applications*, **34**, 2639–2649. 126

VAN DER BURGT, M. (2008). Calibrating low-default portfolios using the cumulative accuracy profile. *Journal of Risk Model Validation*, **1**, 17–33. 103

VAN GESTEL, T. & BAESENS, B. (2009). *Credit Risk Management: Basic Concepts*. Oxford University Press, USA. xxi, 5, 27, 52, 60, 67, 68, 104

VAN GOOL, J., VERBEKE, W., SERCU, P. & BAESENS, B. (2011). Credit scoring for microfinance: is it worth it? *International Journal of Finance & Economics*, **17**, 103–123. 52

VAPNIK, V. (1995). *The Nature of Statistical Learning Theory*. Springer Verlag: New York (USA). 27

VINCIOTTI, V. & HAND, D. (2003). Scorecard construction with unbalanced class sizes. *Journal of Iranian Statistical Society*, **2**, 189–205. 149

WAHLSTRÖM, G. (2009). Risk management versus operational action: Basel II in a Swedish context. *Management Accounting Research*, **20**, 53–68. 63

WANG, M., HUA, X., DAI, L. & SONG, Y. (2006). Enhanced semi-supervised learning for automatic video annotation. In *IEEE International Conference on Multimedia and Expo*, 1485–1488, IEEE. 31

WANG, Y., WANG, S. & LAI, K. (2005). A new fuzzy support vector machine to evaluate credit risk. *Fuzzy Systems, IEEE Transactions on*, **13**, 820–831. 138

WEDEL, M. & KAMAKURA, W. (2000). *Market segmentation: Conceptual and methodological foundations*. Kluwer Academic Publishers: Boston, MA, USA, 2nd edn. 88

WEISS, G. (2004). Mining with rarity: A unifying framework. *Sigkdd Explorations*, **6**, 7–19. 29, 30

WEST, D. (2000). Neural network credit scoring models. *Computers & Operations Research*, **27**, 1131–1152. 5, 7, 105, 113, 126

WESTGAARD, S. & VAN DER WIJST, N. (2001). Default probabilities in a corporate bank portfolio: A logistic model approach. *European Journal of Operational Research*, **135**, 338–349. 105

WILCOXON, F. (1945). Individual comparisons by ranking methods. *Biometrics Bulletin*, **1**, 80–83. 48

WIMS, G., MARTENS, D. & DE BACKER, M. (2011). Network models of financial contagion: A definition and literature review. Tech. Rep. D/2011/7012/35, Faculty of Economics and Business Administration, Ghent University, Belgium, http://feb1.ugent.be/nl/0ndz/wp/Papers/wp_11_730.pdf, last accessed 29 January 2013. 58

WITTEN, I. & FRANK, E. (2005). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann: CA (USA), 2nd edn. 133, 181, 225

WOLFE, D. & HOLLANDER, M. (1999). *Nonparametric statistical methods*. John Wiley: New York (USA), 2nd edn. 177, 178

WORTH, A. & CRONIN, M. (2003). The use of discriminant analysis, logistic regression and classification tree analysis in the development of classification models for human health effects. *Journal of Molecular Structure: THEOCHEM*, **622**, 97–111. 24

WRIGHT, G. (1983). *Building the dream: A social history of housing in America.* MIT Press: MA (USA). 54

XIAO, W., ZHAO, Q. & FEI, Q. (2006). A comparative study of data mining methods in consumer loans credit scoring management. *Journal of Systems Science and Systems Engineering*, **15**, 419–435. 103, 139

XIE, H., HAN, S., SHU, X., YANG, X., QU, X. & ZHENG, S. (2009). Solving credit scoring problem with ensemble learning: A case study. In *Proceedings of the 2nd International Symposium on Knowledge Acquisition and Modeling (KAM'09)*, vol. 1, 51–54, IEEE. 126, 138

XU, P. & CHAN, A. (2003). Support vector machines for multi-class signal classification with unbalanced samples. In *Proceedings of the International Joint Conference on Neural Networks.*, vol. 2, 1116–1119, IEEE. 18

YANG, Z., YOU, W. & JI, G. (2011). Using partial least squares and support vector machines for bankruptcy prediction. *Expert Systems with Applications*, **38**, 8336–8342. 80

YANOVSKIY, K., REVA, E., ZHAVORONKOV, S., SHULGIN, S., LITARCHUK, V., CHERNEY, D., KUCHERINENKO, V. & SHAKIN, D. (2007). Federal reform outcome: Influence of modified institutions on the investment climate in the regions.

Tech. rep., Consortium for Economic Policy Research and Advice, Moscow, available at <http://ssrn.com/abstract=2125182> last accessed 29 January 2013. 80

ZHAO, Y., LI, B., LI, X., LIU, W. & REN, S. (2005). Customer churn prediction using improved one-class support vector machine. *Advanced Data Mining and Applications*, 731–731. 56

ZHU, X. & GOLDBERG, A. (2009). Introduction to semi-supervised learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3, 1–130. 31