UNIVERSITY OF SOUTHAMPTON
FACULTY OF LAW, ARTS & SOCIAL SCIENCES
SCHOOL OF MANAGEMENT

BASEL II COMPLIANT CREDIT RISK MODELLING:

Model development for imbalanced credit scoring data sets, Loss Given Default (LGD) and Exposure At Default (EAD)

Thesis nominated for the degree
of Doctor in Philosophy by:
Iain Leonard Johnston BROWN BBA (Hons) Msc

Supervising committee

Dr. Christophe Mues University of Southampton

Prof. dr. Lyn Thomas University of Southampton

University: UNIVERSITY OF SOUTHAMPTON

ABSTRACT

Faculty: FACULTY OF LAW, ARTS & SOCIAL SCIENCES

School: SCHOOL OF MANAGEMENT

Degree: Doctor of Philosophy

Title: BASEL II COMPLIANT CREDIT RISK MODELLING

Name: By Iain Leonard Johnston Brown

The purpose of this thesis is to determine and to better inform industry practitioners to the most appropriate classification and regression techniques for modelling the three key credit risk components of the Basel II minimum capital requirement; probability of default (PD), loss given default (LGD), and exposure at default (EAD). The Basel II accord regulates risk and capital management requirements to ensure that a bank holds enough capital proportional to the exposed risk of its lending practices. Under the advanced internal ratings based (IRB) approach Basel II allows banks to develop their own empirical models based on historical data for each of PD, LGD and EAD.

In this thesis, first the issue of imbalanced credit scoring data sets, a special case of PD modelling where the number of defaulting observations in a data set is much lower than the number of observations that do not default, is identified, and the suitability of various classification techniques are analysed and presented. As well as using traditional classification techniques this thesis also explores the suitability of gradient boosting, least square support vector machines and random forests as a form of classification. The second part of this thesis focuses on the prediction of LGD, which measures the economic loss, expressed as a percentage of the exposure, in case of default. In this thesis, various state-of-the-art regression techniques to model LGD are considered. In the final part of this thesis we investigate models for predicting the exposure at default (EAD). For off-balance-sheet items (for example credit cards) to calculate the EAD one requires the committed but unused loan amount times a credit conversion factor (CCF). Ordinary least squares (OLS), logistic and cumulative logistic regression models are analysed, as well as an OLS with Beta transformation model, with the main aim of finding the most robust and comprehensible model for the prediction of the CCF. Also a direct estimation of EAD, using an OLS model, will be analysed. All the models built and presented in this thesis have been applied to real-life data sets from major global banking institutions.

Contents

List of Tables and Figures

I. Tables

II. Figures

DECLARATION OF AUTHORSHIP

I, IAIN LEONARD JOHNSTON BROWN,

declare that the thesis entitled:

BASEL II COMPLIANT CREDIT RISK MODELLING,

and the work presented in the thesis are both my own, and have been generated by me as the result of my own original research. I confirm that:

- this work was done wholly or mainly while in candidature for a research degree at this University;
- where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- where I have consulted the published work of others, this is always clearly attributed;
- where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- parts of this work have been published as:

Loterman G, Brown I, Martens D, Mues C, and Baesens B (2012). Benchmarking Regression Algorithms for Loss Given Default Modelling, *International Journal of Forecasting*, **28**(1), 161-170

Brown I & Mues C (2012). An experimental comparison of classification algorithms for imbalanced credit scoring data sets, *Expert Systems with Applications*, **39**(3), Feb 2012, 3446-3453

Signed: ……

Date:……

Acknowledgements

This thesis would not have been possible without the support and guidance of a number of important people, for which I would like to take this opportunity to acknowledge and thank them.

First of all I would like to thank my supervisor, Dr Christophe Mues, who has been of unwavering support throughout my time at the University of Southampton. Without his tutorage and expert knowledge in the field of credit risk modelling I could not have achieved the work conducted in this thesis. I would also like to thank my senior supervisor Prof. Lyn Thomas whose guidance and advice was invaluable in the development of the research topics presented in this thesis.

I have also had the great pleasure of working alongside a number of well established and flourishing academics in the field of credit risk and credit scoring. I would like to thank the team I worked alongside on the LGD project, Gert Loterman, Dr David Martens, Dr Bart Baesens and Dr Christophe Mues. It was a pleasure to work alongside these fine minds in the formulation of our LGD benchmarking paper, which I am happy to say, has now been successfully accepted for publication with the International Journal of Forecasting (IJF). A special thanks also goes to my fellow PhD colleague Gert Loterman for his commitment to the project and presenting our findings together at the Credit Scoring and Credit Control XI Conference in Edinburgh. I wish him all the best in his research and hope to work alongside him again in the future. I would also like to thank everyone who has formed the credit team at the University of Southampton both past and present (Meko, Ed, Mindy, Kasia, Bob, Jie, Ross, Madhur, Angela and Anna). It has been a pleasure to work alongside such friendly, helpful and likeminded persons.

I would like to thank the EPSRC whose financial support over the 3.5 years of this research study allowed me to focus solely on the work at hand. I would also like to take this opportunity to thank SAS UK for their sponsorship and everyone at SAS who I have

met and has helped me over my time as a PhD student. In particular I would like to show my thanks to Geoffrey Taylor who enabled me to partake in the training programmes at the SAS headquarters in Marlow, and for his advice on submitting an application to the SAS Student Ambassador programme which, as a result, I was accepted for. I would also like to thank Dr Laurie Miles for his support and interest in the work and papers that have been completed as part of this thesis.

I have also worked alongside and come into contact with a number of amazing people during my time at the University of Southampton. In particular I would like to thank Gillian Groom not only for her help in enabling me to teach SAS but also for the part she played in facilitating my future career, for this I am indebted to her. A special thanks goes to my colleague and flatmate Mindy (and her husband Nic) for their support and Wii parties throughout my time in Southampton. The friendship of both of them made life a lot more interesting and enjoyable. I would also like to thank Shivam and Joe, not only for being fellow PhD colleagues in the School of Management but for their football skills on the pitch and being part of the mighty Lazy Town FC.

Finally I would like to express my greatest thanks to my whole family, who without their support I would not have achieved any of the goals I have set out to attain. I owe a huge debt (both monetarily and emotionally) to my parents and sister who have provided me with the love and means necessary to succeed and achieve where I am today. I would also like to thank my fiancé who has supported me throughout my PhD and whenever I have felt disheartened with my work her love, encouragement and incredible cooking has put me back on track.

Chapter 1

1 Introduction

With the recent financial instabilities in the credit markets, the area of credit risk modelling has become ever more important, leading to the need for more accurate and robust models. Further to this, the introduction of the Basel II Capital Accord (Basel Committee on Banking Supervision, 2004) now allows for financial institutions to derive their own internal credit risk models under the advanced internal ratings based approach (AIRB). The Basel II Capital Accord prescribes the minimum amount of regulatory capital an institution must hold so as to provide a safety cushion against unexpected losses. From a credit risk perspective, and under the advanced internal ratings based approach (AIRB), the accord allows financial institutions to build risk models for three key risk parameters: Probability of Default (PD), Loss Given Default (LGD), and Exposure at Default (EAD). The Probability of Default (PD) is defined as the likelihood that a loan will not be repaid and will therefore fall into default. Loss Given Default (LGD) is the estimated economic loss, expressed as a percentage of exposure, which will be incurred if an obligor goes into default. Exposure at Default (EAD) is a measure of the monetary exposure should an obligor go into default.

In this thesis, we study the use of classification and regression techniques to develop models for the prediction of all three components of expected loss, Probability of Default (PD), Loss Given Default (LGD) and Exposure At Default (EAD). The reason why these particular topics have been chosen is due in part to the increased scrutiny on the financial sector and the pressure on them by the financial regulators to move to and advanced

internal ratings based approach. The financial sector is therefore looking to the best models possible to determine their minimum capital requirements through the estimation of PD, LGD and EAD. On the issue of PD estimation a great deal of work has already been conducted in both academia and the industry; therefore in Chapter 4 we will tackle a special case of PD modelling, i.e. building default prediction models for imbalanced credit scoring data sets. In a credit scoring context imbalanced data sets frequently occur when the number of defaulting loans in a data set is much lower than the number of observations that do not default. Subsequently, in Chapters 5 and 6, we then turn our attention to the other two much less researched risk components of LGD and EAD. It is our aim to validate novel approaches, evaluate their effectiveness for all three components of expected loss and obtain an improved understanding of the risk drivers in the prediction of EAD.

This introduction chapter is structured as follows. We will begin by giving a detailed background of the Basel II Capital Accord, with emphasis on its implications to credit risk modelling. We will then go on to introduce the three extensive projects which tackle the issues highlighted in PD, LGD and EAD modelling as well as the motivations for choosing these research topics. A list of contributions will also be given after the introduction of each of the projects. Finally a brief description of the notations used throughout this thesis will be detailed.

1.1 The Basel II Capital Accord

The banking/financial sector is one of the most closely scrutinised and regulated industries and as such subject to stringent controls. The reason for this is that banks can only lend out money in the form of loans if depositors trust that the bank and the banking system is stable enough and their money will be there when they require to withdraw it. However, in order for the banking sector to provide the loans and mortgages they must leverage depositors' savings meaning that only with this trust can they continue to function. It is imperative therefore to prevent a loss of confidence and distrust in the

banking sector from occurring, as it can have serious implications to the wider economy as a whole.

The job of the regulatory bodies therefore is to contribute to ensuring the necessary trust and stability by limiting the level of risk that banks are allowed to take. In order for this to effectively work, the maximum risk level banks can take needs to be set in relation to the bank's own capital. From the banks perspective the high cost of acquiring and holding capital makes it prohibitive and unfeasible to have it fully cover all of a bank's risks. As a compromise, the major regulatory body of the banking industry, the Basel Committee on Banking Supervision, proposed guidelines in 1988 whereby a solvability coefficient of eight percent was introduced, i.e. the total assets, weighted for their risk, must not exceed eight percent of the bank's own capital, Basel I (SAS Institute, 2002).

The figure of eight percent assigned by the Basel Committee is somewhat arbitrary and as such since the conception of the idea has been subject to much debate. After the introduction of the Basel I accord more than one hundred countries worldwide adopted the guidelines, becoming a major milestone in the history of global banking regulation. However, a number of the accord's inadequacies, in particular with regard to the way that credit risk was measured, became apparent over time (SAS Institute, 2002). To account for these issues a revised accord, Basel II, was conceived.

As defined the Basel II Capital Accord (Basel Committee on Banking Supervision, 2001a) prescribes the minimum amount of regulatory capital an institution must hold so as to provide a safety cushion against unexpected losses. The Accord comprises of three pillars:

Pillar 1: Minimum Capital Requirements
Pillar 2: Supervisory Review Process
Pillar 3: Market Discipline and Public Disclosure

Pillar 1 aligns the minimum capital requirements to a bank's actual risk of economic loss. Various approaches to calculating this are prescribed in the accord (including more risk-

sensitive standardized and internal ratings-based approaches) which will be described in more detail. Pillar 2 entails supervisors evaluating the activities and risk profiles of banks to determine whether they should hold higher levels of capital than those prescribed by Pillar 1 and offers guidelines for the supervisory review process, including the approval of internal rating systems. Pillar 3 leverages the ability of market discipline to motivate prudent management by enhancing the degree of transparency in banks' public disclosure. (Basel, 2004).

The Basel II Capital Accord entitles banks to compute their credit risk capital in either of two ways:

1. Standardised Approach
2. Internal Ratings Based (IRB) Approach
    a. Foundation Approach
    b. Advanced Approach

Under the standardised approach banks are required to use ratings from external credit rating agencies to quantify required capital. The main purpose and strategy of the Basel committee is to offer capital incentives to banks that move from a supervisory approach to a best practice advanced internal ratings based one. The two versions of the internal ratings based (IRB) approach permit banks to develop and use their own internal risk ratings, to varying degrees. The IRB approach is based on the following four key parameters:

1. Probability of Default (PD): the likelihood that a loan will not be repaid and will therefore fall into default;
2. Loss Given Default (LGD): the estimated economic loss, expressed as a percentage of exposure, which will be incurred if an obligor goes into default, in other words, LGD equals: 1 minus the recovery rate;
3. Exposure At Default (EAD): a measure of the monetary exposure should an obligor go into default;

4. Maturity (M): is the length of time to the final payment date of a loan or other financial instrument.

From the parameters, PD, LGD and EAD, expected loss (EL) can be derived as follows:

$$EL = PD \times LGD \times EAD. \quad (1.1)$$

For example, if $PD = 2\%, LGD = 40\%, EAD = £10,000$, then $EL = £80$. Expected loss can also be measured as a percentage of EAD:

$$EL\% = PD \times LGD. \quad (1.2)$$

In the previous example expected loss as a percentage of EAD would be equal to $EL\% = 0.8\%$. The internal rating based approach requires financial institutions to estimate values for PD, LGD and EAD for their various portfolios. Two IRB options are available to financial institutions; a foundation approach and an advanced approach (FIGURE 1.1) (Basel Committee on Banking Supervision, 2001a):

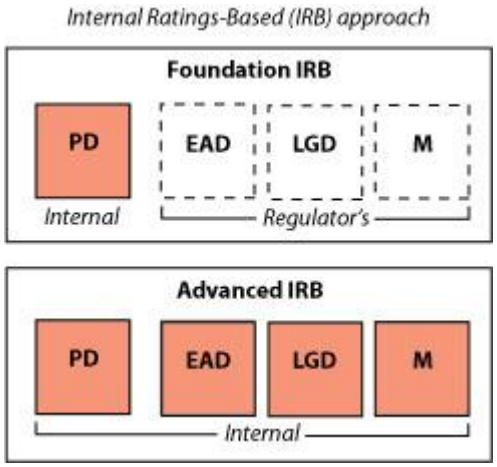


FIGURE 1.1: Illustration of foundation and advanced Internal Ratings-Based (IRB) approach

The difference between these two approaches is the degree to which the four parameters can be measured internally. For the foundation approach, only PD may be calculated internally, subject to supervisory review (Pillar 2). The values for LGD and EAD are

fixed and based on supervisory values. For the final parameter, M, a single average maturity of 2.5 years is assumed for the portfolio. In the advanced IRB approach all four parameters are to be calculated by the bank and are subject to supervisory review (Schuermann, 2004).

Under the AIRB, financial institutions are also recommended to estimate a 'Downturn LGD', which 'cannot be less than the long-run default-weighted average LGD calculated based on the average economic loss of all observed defaults with the data source for that type of facility' (Basel, 2004).

We will now look at and identify some of the problems faced by financial institutions wishing to implement the advanced IRB approach.

1.2 The imbalanced credit scoring data set problem (a special case of probability of default (PD) modelling)

Commonly the first stage of PD estimation involves building a scoring model that can be used to distinguish between different risk classes. In the development of credit scoring models several statistical methods are used traditionally such as linear probability models, logit models and discriminate analysis models. These statistical techniques can be used to estimate the probability of default of a borrower based on factors such as loan performance and the borrowers' characteristics. Based on this information credit scorecards can be built to determine whether to accept or decline a borrower (application scoring) or to provide an up-to-date assessment of the credit risk of existing borrowers (behavioural scoring). The aim of credit scoring therefore is essentially to classify loan applicants into two classes, i.e. good payers (i.e., those who are likely to keep up with their repayments) and bad payers (i.e., those who are likely to default on their loans) (Thomas, 2000) . In the current financial climate, and with the recent introduction of the Basel II Accord, financial institutions have even more incentives to select and implement the most appropriate credit scoring techniques for their credit data sets. It is stated in

Henley and Hand (1997) that companies could make significant future savings if an improvement of only a fraction of a percent could be made in the accuracy of the credit scoring techniques implemented. However, in the literature, data sets that can be considered as very low risk, or imbalanced data sets, have had relatively little attention paid to them in particular with regards to which techniques are most appropriate for scoring them (Benjamin *et al*, 2006). The underlying problem with imbalanced data sets is that they contain a much smaller number of observations in the class of defaulters than in that of the good payers. A large class imbalance is therefore present which some techniques may not be able to successfully handle (Benjamin *et al*, 2006). In a recent FSA publication regarding conservative estimation of imbalanced data sets, regulatory concerns were raised about whether firms can adequately asses the risk of imbalanced credit scoring data sets (Benjamin *et al*, 2006).

A wide range of classification techniques have already been proposed in the credit scoring literature, including statistical techniques, such as linear discriminant analysis and logistic regression, and non-parametric models, such as k-nearest neighbour and decision trees. But it is currently unclear from the literature which technique is the most appropriate for improving discrimination for imbalanced credit scoring data sets. TABLE 1.1 in Section 2.1 provides a selection of techniques currently applied in a credit scoring context, not specifically for imbalanced data sets, along with references showing some of their reported applications in the literature.

Hence, the aim of the first project, reported in Chapter 4, is to conduct a study of various classification techniques based on five real-life credit scoring data sets. These data sets will then have the size of their minority class of defaulters further reduced by decrements of 5% (from an original 70/30 good/bad split) to see how the performance of the various classification techniques is affected by increasing class imbalance.

The five real-life credit scoring data sets to be used in this empirical study include two data sets from Benelux (Belgium, Netherlands and Luxembourg) institutions, the German Credit and Australian Credit data sets which are publicly available at the UCI repository (http://kdd.ics.uci.edu/), and the fifth data set is a behavioural scoring data set, which was also obtained from a Benelux institution.

The techniques that will be evaluated in this chapter are traditional well reported classification techniques (Baesens, *et al* 2003); logistic regression (LOG), linear and quadratic discriminant analysis (LDA, QDA), nearest-neighbour classifiers (k-NN10, k-NN100), decision trees (C4.5) and more machine learning techniques; least square support vector machines (LS-SVM), neural networks (NN), a gradient boosting algorithm and random forests. The reason why these machine learning techniques are selected are their potential applications in a credit scoring context (Baesens, *et al* 2003) and the interest in whether they can perform better than traditional techniques given a large class imbalance. We are especially interested in the power and usefulness of the gradient boosting and random forest classifiers which have yet to be thoroughly investigated in a credit scoring context.
All techniques will be evaluated in terms of their Area Under the Receiver Operating Characteristic Curve (AUC). This is a measure of the discrimination power of a classifier without regard to class distribution or misclassification cost (Baesens, *et al* 2003).
To make statistical inferences from the observed differences in AUC, we will follow the recommendations given in a recent article, Demšar 2006, which looked at the problem of benchmarking classifiers on multiple data sets and recommended a set of simple robust non-parametric tests for the statistical comparison of the classifiers. The AUC measures will therefore be compared using Friedman's average rank test, and Nemenyi's post-hoc test will be used to test the significance of the differences in rank between individual classifiers. Finally, a variant of Demšar's significance diagrams will be plotted to visualise their results.

Having introduced the topic of research which will be conducted in Chapter 4 we will now identify the major motivations for this thesis chapter.

Fundamentally from a regulatory perspective the issue is whether firms can adequately build loan-level scoring models on imbalanced data sets as not all techniques may be able to cope well with class imbalances; as a result, discrimination performance may suffer. Without an adequate scoring model, it becomes difficult to segment exposures into

different rating grades or pools. So the key question becomes not whether we can assess the risk but can we still build a decent model that distinguishes between different levels of (low) risk. Thus the topic in this research thesis has been chosen so as to assess the capabilities of credit scoring techniques when a large class imbalance is present. The motivation behind this particular research topic is to identify the capabilities of traditional techniques such as logistic regression and linear discriminant analysis when a class imbalance is present and compare these to techniques yet to be analysed in this field i.e. gradient boosting and random forests. If for example logistic regression can perform comparatively well to the more advanced techniques, when a large class imbalance is present, this will provide confidence to practitioners wishing to implement such a technique.

The experimental design has been chosen so that a variety of available datasets can be compared at varying levels of class imbalance (through under sampling the bad observations). A process of 10-fold cross validation is applied to retain statistical and empirical inference where small numbers of bad observations are present in the imbalanced samples. Further motivations behind the experimental design of this particular research area are identified and assessed in the literature review section of this thesis.

1.3 The estimation of Loss Given Default (LGD)

The LGD parameter measures the economic loss, expressed as percentage of the exposure, in case of default. This parameter is a crucial input to the Basel II capital calculation as it enters the capital requirement formula in a linear way (unlike PD, which comparatively has a smaller effect on minimal capital requirements). Hence, changes in LGD directly affect the capital of a financial institution and as such also its long-term strategy. It is thus of crucial importance to have models that estimate LGD as accurately as possible. This is however not straightforward, as industry models typically show low R^2 values. Such models are often built using ordinary least squares regression or

regression trees, even though prior research has shown that LGD typically displays a non-linear bi-modal distribution with spikes around 0 and 1 (Bellotti & Crook 2007). In the literature the majority of work to date has focused on the issues related to PD estimation whereas only more recently, academic work has been conducted into the estimation of LGD (e.g. Bellotti and Crook, 2009, Loterman *et al*, 2009, Thomas *et al*, 2010).
In Chapter 5, a large set of state-of-the-art regression algorithms will be applied to 6 real-life LGD data sets with the aim of achieving a better understanding of which techniques perform the best in the prediction of LGD. The regression models employed will include one-stage models, such as those built by ordinary least squares, beta regression, artificial neural networks, support vector machines and regression trees, as well as two-stage models which attempt to combine the benefits of multiple techniques. Their performances will be determined through the calculation of several performance metrics which will in turn be meta-ranked to determine the most predictive regression algorithm. The performance metrics will again be compared using Friedman's average rank test and Nemenyi's post-hoc test will be employed to test the significance of the differences in rank between individual regression algorithms. Finally, a variant of Demšar's significance diagrams will be plotted for each performance metric to visualise their results.
This first large scale LGD benchmarking study in terms of techniques and data sets, investigates whether other approaches can improve the predictive performance which, given the impact of LGD on capital requirements, can yield large benefits.

Having introduced the topic of research which will be conducted in Chapter 5 we will now identify the major motivations for this thesis chapter.

There has been much industry debate as to the best techniques to apply in the estimation of LGD, given its bi-modal distribution. The motivations for this particular research topic are to determine the predictive power of commonly used techniques such as linear regression with transformations and compare them to more advanced machine learning techniques such as neural networks and support vector machines. The aim in doing this is to better inform industry practitioners as to the comparable ability of potential techniques

and to add to the current literature on both the topics of loss given default and applications of domain specific regression algorithms.

1.4 Model development for Exposure At Default (EAD)

Over the last few decades, credit risk research has largely been focused on the estimation and validation of probability of default (PD) models in credit scoring. However, to date very little model development and validation has been reported on the estimation of EAD, particularly for retail lending (credit cards). As with LGD, EAD enters the capital requirement formulas in a linear way and therefore changes to EAD estimations have a crucial impact on regulatory capital. Hence, as with LGD, it is important to develop robust models that estimate EAD as accurately as possible.

In defining EAD for on-balance sheet items, EAD is typically taken to be the nominal outstanding balance net of any specific provisions (Financial Supervision Authority, UK 2004a, 2004b). For off-balance sheet items (for example, credit cards), EAD is estimated as the current drawn amount, $E(t_r)$, plus the current undrawn amount (i.e. credit limit minus drawn amount), $L(t_r) - E(t_r)$, multiplied by a credit conversion factor, CCF or loan equivalency factor (LEQ):

$$EAD = E(t_r) + CCF.\left(L(t_r) - E(t_r)\right). \tag{1.3}$$

The credit conversion factor can be defined as the percentage rate of undrawn credit lines (UCL) that have yet to be paid out but will be utilised by the borrower by the time the default occurs (Gruber and Parchert, 2006). The calculation of the CCF is very important for off-balance sheet items as the current exposure is generally not a good indication of the final EAD, the reason being that, as an exposure moves towards default, the likelihood is that more will be drawn down on the account. In other words, the source of variability of the exposure is the possibility of additional withdrawals when the limit allows this (Moral, 2006).

The purpose of this chapter will therefore be to look at the estimation and validation of this CCF in order to correctly estimate the off-balance sheet EAD. A real-life data set with monthly balance amounts for clients over the period 2001-2004 will be used in the building and testing of the regression models. We also aim to gain a better understanding of the variables that drive the prediction of the CCF for consumer credit. To achieve this, predictive variables that have previously been suggested in the literature (Moral, 2006) will be constructed, along with a combination of new and potentially significant variables. We also aim to identify whether an improvement in predictive power can be achieved over ordinary least squares regression (OLS) by the use of binary logit and cumulative logit regression models and an OLS with Beta transformation model. The reason why we propose these models is that recent studies (e.g. Jacobs, 2008) have shown that the CCF exhibits a bi-modal distribution with two peaks around 0 and 1, and a relatively flat distribution between those peaks. This non-normal distribution is therefore less suitable for modelling with traditional ordinary least squares (OLS) regression. The motivation for using an OLS with Beta transformation model is that it accounts for a range of distributions including a U-shaped distribution. We will also trial a direct OLS estimation of the EAD and use it as a comparison to estimating a CCF and applying it to the EAD formulation.

Having introduced the topic of research which will be conducted in Chapter 6 we will now identify the major motivations for this thesis chapter.

The correct calculation of credit conversion factors for off-balance sheet items is of pertinent interest and importance to the financial sector. The main motivation for choosing this research topic therefore is to provide insight to the industry as to the potential techniques at their disposal for calculation their CCFs. The estimation of CCF is also a similar problem to that of estimating LGD, given that it displays a bi-modal distribution.

The experimental design was chosen to assess a variety of techniques on a revolving credit data set and compare their predictive power as well as their ability to provide robust results for the actual exposure at default estimation.

1.5 Contributions

Having identified the need for a greater understanding of the appropriate credit risk modelling techniques available to practitioners, we will now identify the major research topics and contributions of this thesis chapter.

1.5.1 Building default prediction models for imbalanced credit scoring data sets

The contributions of the research set out in Chapter 4 of this thesis are as follows.
In Chapter 4 of this thesis, we will address this issue of estimating probability of default for imbalanced data sets. Whereas other studies have benchmarked several scoring techniques, in our study, we have explicitly looked at the problem of having to build models on potentially highly imbalanced data sets. Two techniques that have yet to be fully researched in the context of credit scoring, i.e. Gradient Boosting and Random Forests, will be chosen, alongside traditional credit scoring techniques, to give a broader review of the techniques available.
The results of these experiments will show that the Gradient Boosting and Random Forest classifiers perform well in dealing with samples where a large class imbalance is present. The findings will also suggest that the use of a linear kernel LS-SVM is not beneficial in the scoring of data sets where a very large class imbalance exists.

1.5.2 Estimation of Loss Given Default (LGD)

In Chapter 5, a large scale Loss Given Default (LGD) benchmarking study will be undertaken, with the aim of comparing various state-of-the-art regression techniques to model and predict LGD. The findings displayed in Chapter 5 will indicate that the average predictive performance of the models in terms of R^2 ranges from 4 % to 43 %, indicating that most resulting models have limited explanatory power. Nonetheless, a clear trend will be displayed showing that non-linear techniques and artificial neural networks and support vector machines in particular give higher performances than more traditional linear techniques. This indicates the presence of non-linear interactions between the independent variables and the LGD, contrary to some studies in PD modelling where the difference between linear and non-linear techniques is not that explicit. Given the fact that LGD has a bigger impact on the minimal capital requirements than PD, we will demonstrate the potential importance of applying non-linear techniques, preferably in a two-stage context to obtain comprehensibility as well, for LGD modelling. To the best of our knowledge, such an LGD study has not yet been conducted before in the literature.

1.5.3 Regression model development for Credit Card Exposure At Default (EAD)

In Chapter 6, we will propose several models for predicting the Exposure At Default (EAD) and estimating the credit conversion factor (CCF). Ordinary least squares, binary logit and cumulative logit regression models will be estimated and compared for the prediction of the CCF, which to date have not been thoroughly evaluated before. A variety of new variables of interest will also be calculated and used in the prediction of the CCF. An in-depth analysis of the predictive variables used in the modelling of the CCF will be given, and will show that previously acknowledged variables are significant. The results from this chapter will also show that a marginal improvement in the coefficient of determination can be achieved with the use of a binary logit model over a

traditional OLS model. Interestingly the use of a cumulative logit model is shown to perform worse than both the binary logit and OLS models.

With regards to the additional variables proposed in the prediction of the CCF, only one, i.e. average number of days delinquent in the last 6 months, gives an adequate p-value when a stepwise procedure was used.

1.6 Notation

In this thesis, the following mathematical notations are used. A scalar x is denoted in normal script. A vector $\mathbf{x}$ is represented in boldface and is assumed to be a column vector, $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}$. The corresponding row vector $\mathbf{x}^{\mathbf{T}}$ is obtained using the transpose T, $\mathbf{x}^{\mathbf{T}} = \begin{bmatrix} x_1 \\ x_2 \\ \dots \\ x_n \end{bmatrix}^T = \begin{bmatrix} x_1 & x_2 & \dots & x_n \end{bmatrix}$. Bold capital notation is used for a matrix, $\mathbf{X}$. The number of independent variables is given by n and the number of observations is given by l. The observation i is denoted as $\mathbf{x}_i$ whereas variable j is indicated as x_j. The value of variable j for observation i is represented as $x_i(j)$ and the dependent variable y for observation i is represented as y_i. P is used to denote a probability. p is used to denote a proportion.

Chapter 2

2 Literature Review

In this section a review of the literature topics related to this PhD thesis will be given. This section is formulated as follows. We begin by looking at the current applications of data mining techniques in credit risk modelling and go on to look at the current work and issues in the modelling of the three parameters of the minimum capital requirements (probability of default, loss given default and exposure at default). To date a considerable amount of work has been done on the estimation of the probability of default. To further this, the issue of imbalanced credit scoring data sets, which has been highlighted by the Basel Committee on Banking Supervision (2005) as a potential problem for probability of default modelling, is also looked at and reviewed. Finally, a summary of the literature review chapter will be given.

2.1 Current applications of data mining techniques in credit risk modelling

In this section, a review of the current applications of data mining techniques in a credit risk modelling environment will be given. The ideas already present in the literature will be explored with the aim to highlight potential gaps with which further research could fill. TABLE 1.1 provides a selection of techniques currently applied in a credit scoring context, not specifically for imbalanced data sets, along with references showing some of their reported applications in the literature.

Classification Techniques	Application in a credit scoring context
Logistic Regression (LOG)	Arminger, *et al* (1997), Baesens, *et al* (2003), Desai, *et al* (1996), Steenackers & Goovaerts (1989), West (2000), Wiginton (1980)
Decision Trees (C4.5, CART, etc.)	Arminger, *et al* (1997), Baesens, *et al* (2003), West (2000), Yobas, *et al* (2000)
Neural Networks (NN)	Altman (1994), Arminger, *et al* (1997), Baesens, *et al* (2003), Desai, *et al* (1996), West (2000), Yobas, *et al* (2000)
Linear Discriminant Analysis (LDA)	Altman (1968), Baesens, *et al* (2003), Desai, *et al* (1996), West (2000), Yobas, *et al* (2000)
Quadratic Discriminant Analysis (QDA)	Altman (1968), Baesens, *et al* (2003)
k-Nearest Neighbours (k-NN)	Baesens, *et al* (2003), Chatterjee & Barcun (1970), West (2000)

Support Vector Machines (SVM, LS-SVM, etc.)	Baesens, *et al* (2003), Huang (2007), Yang (2007)

TABLE 1.1: Credit scoring techniques and their applications

In the development of credit risk modelling and scorecard building, discriminant analysis and linear or logistic regression have traditionally been the most widely applied techniques (Hand & Henley, 1997). This is partly due to their ability to be easily understood and ease of application. The first major work in the application of machine-learning algorithms in a credit risk modelling context was conducted by Davis et al (1992). In this paper a number of algorithms, including Bayesian inference and neural networks, are applied to credit-card assessment data from the Bank of Scotland. Their findings suggest that overall all the algorithms analysed perform at the same level of accuracy, with the neural network algorithms taking the longest to train. Their research was limited however by the number of data observations in both the training and test sets, and by the computational power of the period. Further research has since been conducted into the applications of data mining techniques over a larger selection of data sets however, and the findings from these studies will be discussed before conclusions to potential gaps are made.

To date, a variety of data mining models have been used in the estimation of default for both consumer and commercial credit. In Rosenberg and Gleit (1994), a survey of the use of discriminant analysis, decision trees, expert systems for static decisions, dynamic and linear programming and Markov chains is undertaken for credit management. They surmised that although, up until that period, sophisticated techniques such as linear or dynamic programming were unused in practice, there was a potential future use for them in this context. This signified the potential for other practitioners to further their study and apply techniques such as linear programming in the estimation of credit risk.
Hand and Henley (1997) examined the problems that have arisen in the credit scoring context as well as giving a detailed review of the statistical methods used. They state that although the main focus of statistical methods for credit scoring has so far been to simply discriminate between good and bad risk classes, there is a much larger scope for the application of these techniques. This leads to the application of data mining techniques in

a credit risk modelling environment, such as modelling risk parameters for Basel. Further discussion of the implications and practicalities of applying these methods in a credit risk modelling domain will follow the discussion of additional credit scoring techniques and research. Similarly to Hand and Henley (1997) in Lee et al (2002) widely used techniques (such as logistic regression and linear discriminant analysis) are compared as well as exploring the integration of back-propagation neural networks are with traditional discriminant analysis with the aim of improving credit scoring performance. Their findings indicate that not only can convergence be achieved quicker than with neural networks on their own, but in terms of accuracy an improvement over logistic regression and discriminant analysis can be made.

Expanding on the work conducted by Davis et al (1992), Giudici (2001) identifies the use of, Bayesian methods, coupled with Markov Chain Monte Carlo (MCMC) computational techniques are shown to be successfully employed in the analysis of highly dimensional complex data sets, as are common in data mining. This study shows the potential of MCMC for credit scoring. Through the use of a reversible jump MCMC and graphical models, one can extract valuable information from data sets, in the form of conditional dependence relationships. Applications of MCMC in the specific context of modelling LGD (discussed in section 2.2.2) can also be found in Luo & Shevchenko (2010).

In Baesens et al (2003), the performance of various state-of-the-art classification algorithms are applied to eight real-life credit scoring data sets. Their findings suggest that while simple classifiers such as logistic regression and linear discriminant analysis yield reasonable results in a credit scoring context, more novel techniques such as LS-SVMs and neural networks can yield improved results. Their findings also indicate that as the traditional linear techniques provided reasonable performance, credit scoring data sets are only weakly non-linear. The work presented in thesis attempts to build upon the findings shown in this paper, but for the case where class imbalances are present. This will test the hypotheses put forward in Baesens et al (2003) when similar classification techniques are applied over varying levels of class imbalance. There is a clear possibility for future research, shown in this paper, over a wider range of credit data sets and through using a wider breath of machine learning techniques.

An extension of Baesens et al (2003) can be seen in Van Gestel et al (2005) where a combination approach to credit scoring is adopted through the implementation of both linear (logistic regression) and non-linear (support vector machines) techniques. It is shown that through a gradual approach of combining the readability of logistic regression and the complexity of support vector machines improved accuracy of performance is observed. The use of SVM's allows the combined model to capture multivariate non-linear relations. This study will form the basis of the extended work in Chapter 5 of this thesis where we look to expand on this potential modelling process through the use of other non-linear techniques in combination with linear ones.

More recently a comparison of a variety of data mining techniques is given in Yeh and Lien (2009). In their paper the predictive accuracy of six data mining methods are compared (K-nearest neighbours, Logistic Regression, Discriminant Analysis, Naïve Bayesian classifiers, Artificial Neural Networks and Classification Trees) on customers' default payments in Taiwan. For this paper the predictive accuracy of the estimated probability of default is analysed as opposed to a traditional classification analysis. The findings indicate that the forecasting model produced by artificial neural networks has the highest coefficient of determination (R-Square) in estimating the real probability of default. This goes someway in agreeing with the findings shown in Baesens et al. (2003) and hence strengths the need to identify how well these techniques can still perform give varying levels of class imbalance in a credit scoring context.

It must be noted that this literature overview for current data mining techniques applied in a credit risk modelling context is by no means exhaustive. Other techniques used for credit scoring and risk modelling include for example, genetic algorithms (Bedingfield and Smith, 2001, Fogarty et al. 1992) and mathematical programming (Freed and Glover, 1981, Hand and Jacka, 1981, Kolesar and Showers, 1985).

The majority of the studies reviewed here display the same limitations in numbers of real-world credit data sets used and number/variety of techniques compared. Another consideration is the fact that the area under the receiver operator curve (AUC) statistic is

not widely reported in these studies, whereas in industry practice this is a well understood and well used statistical measure. This thesis therefore will attempt to incorporate a wide variety of techniques and real world credit data sets and provide performance metrics that are of use within industry practice (i.e. R-square for regression models, AUC for classification models and correlation metrics).

For a more detailed review paper of the statistical classification methods used in consumer credit scoring please see Hand and Henley (1997).

2.2 Components

This section details the literature studies on the three contributing components to the calculation of the minimum capital requirements. The current understanding and implementations of these in the literature will be discussed.

2.2.1 Probability of Default (PD)

Over the last few decades, the main focus of credit risk modelling literature has focused on the estimation of the probability of default on individual loans or pools of transactions (PD); with less literature available on the estimation of the loss given default (LGD) and the correlation between defaults (Crouhy et al, 2000; Duffie & Singleton, 2003). Work has also been developed on exposure at default modelling, but to a far lesser extent (cf. section 2.2.3).

Probability of default (PD) can be defined as the likelihood that a loan will not be repaid and will therefore fall into default. A default is considered to have occurred with regard to a particular obligor (i.e. customer) when either or both of the two following events have taken place:

1. The bank considers that the obligor is unlikely to pay its credit obligations to the banking group in full (e.g. if an obligor declares bankruptcy), without recourse by the bank to actions such as realising security (if held) (i.e. taking ownership of the obligors house, if they were to default on a mortgage).
2. The obligor is past due, i.e. missed payments, for more than 90 days on any material credit obligation to the banking group. (Basel, 2004)

This section gives a non-extensive overview of the key literature to date in the field of PD modelling. A clear distinction can be made between those models developed for retail credit and corporate credit facilities in the estimation of PD. As such this section has been sub-divided into three categories distinguishing the literature for retail credit (cf. 2.2.1.a), corporate credit (cf. 2.2.1.b) and calibration (cf. 2.2.1.c).

2.2.1.a PD models for retail credit

Credit scoring analysis is the most well known and widely used model to measure default risk in consumer lending. Historically most credit scoring models are based on the use of historical loan and borrower data to identify which characteristics are able to distinguish between defaulted and non-defaulted loans (Giambona & Iancono, 2008). Other detailed references of the credit scoring literature can be found in Mays (1998), Hand and Henley (1997), Mester (1997), Viganò (1993), and Lewis (1990). These papers provide a variety of applications in modelling PD and are mentioned here as a pointer to a further review of the current literature. Hand and Henley (1997) is discussed in more detail in a prior section 2.1. In terms of the credit scoring models used in practice, the following list highlights the five main traditional forms:

(1) Linear probability models (Altman, 1968);
(2) Logit models (Martin, 1977);
(3) Probit models (Ohlson, 1980);
(4) Multiple discriminant analysis models and,
(5) Decision trees.

(Giambona & Iancono, 2008)

The main benefits of credit scoring models are their relative ease of implementation and the fact that they do not suffer from the opaqueness of some of the other proposed "black-box" techniques such as Neural Networks and Least Square Support Vector Machines proposed in Baesens et al (2003).

Since the advent of the new capital accord (Basel Committee on Banking Supervision, 2004), a renewed interest has been seen in credit risk modelling. With the allowance under the internal ratings based approach of the capital accord for organisations to create their own internal ratings models, the use of appropriate modelling techniques is ever more prevalent. Banks must now weigh up the issue of holding enough capital to limit

insolvency risks and not holding excessive capital due to its cost and limits to efficiency (Bonfim, 2009).

Further recent work on the discussion of PD estimation from a regulatory perspective for retail credit can be found in Chatterjee et al (2007), where the consequences of changes in regulation of bankruptcy are analysed and advisory pointers given.

2.2.1.b PD models for corporate credit

With regards to corporate PD models, Crouhy et al. (2000) identify the more recent contributions to the field of PD modelling identifying the concepts behind the KMV RiskCalc and CreditPortfolioView models. The KMV RiskCalc model adopts a microeconomic approach relating the probability of default of any obligor to the market value of its assets. CreditPortfolioView however takes into account macroeconomic factors to default and migration probabilities. Similarly, Gordy (2000) offers a comparative anatomy of credit risk models, including RiskMetrics Group's CreditMetrics and Credit Suisse Financial Product's CreditRisk+. It is shown that although these are comparatively different packages the underlying mathematical structures are very similar. Simulation exercises are also run to evaluate the effects of each of the differences in the packages. Further to this Murphy et al (2002) provide an application of a RiskCalc model on private Portugese firms.

With regards to benchmarking classification techniques on corporate credit data, West (2000) provides a comprehensive study of the credit scoring accuracy of five neural network models on two corporate credit data sets. The neural network models are then benchmarked against traditional techniques such as linear discriminant analysis, logistic regression and k-nearest neighbours. The findings demonstrate that although the neural network models perform well more simplistic, logistic regression is a good alternative with benefit of being much more readable and understandable. A limiting factor of this study is it only focuses on the application of additional neural network techniques on two relatively small data sets, and doesn't take into account larger data sets or other machine learning approaches. The topic of research presented in this thesis aims to extend the work conducted by West (2000) into the arena of additional machine learning techniques

and to also test the capabilities of these techniques when class imbalances are present. Other recent work on PD estimation for corporate credit can be found in Fernandes (2005), Carling et al (2007), Tarashev (2008), Miyake and Inoue (2009) and Kiefer (2010).

2.2.1.c PD calibration

The purpose of PD calibration is the assignment of a default probability to each possible score or rating grade values. The important information required for calibrating PD models include:

- The PD forecasts over a rating class and the credit portfolio for a specific forecasting period.
- The number of obligors assigned to the respective rating class by the model.
- The default status of the debtors at the end of the forecasting period.

(Guettler and Liedtke, 2007)

It has been found (Guettler and Liedtke, 2007) that realised default rates are actually subject to relatively large fluctuations making it necessary to develop indicators to show how well a rating model estimates the PDs. It is recommended in Tasche (2003), that traffic light indicators could be used to show whether there is any significance in the deviations of the realised and forecasted default rates. The three traffic light indicators, green, yellow and red identify the following potential issues. A green traffic light indicates that the true default rate is equal to, or lower than, the upper bound default rate at a low confidence level. A yellow traffic light indicates the true default rate is higher than the upper bound default rate at a low confidence level and equal to, or lower than, the upper bound default rate at a high confidence level. Finally a red traffic light indicates the true default rate is higher than the upper bound default rate at a high confidence level. (Tasche, 2003 via Guettler and Liedtke, 2007)

Although a non-exhaustive list, substantial work has previously been conducted in the estimation of probability of default. This section of literature is included to inform the

reader of the current modelling research to date with regards to PD, which will form a precursor to the analysis of credit scoring for imbalanced data sets. As the topic of research of this thesis is focused towards estimating PD in imbalanced datasets a more exhaustive review of the current literature on Probability of Default modelling can be found in the following review papers; Altman and Sironi (2004), Erdem C (2008). However, as we will see in the next section, an interesting finding is that little work has been conducted on the area of imbalanced data sets, where there are a much smaller number of observations in the class of defaulters than in that of the class of payers, where a PD estimate must also be achieved. Therefore in the following section, the issue of imbalanced credit scoring data sets will be looked at with the aim to identify the current approaches in the literature and identify any potential gaps.

2.2.1.1 Imbalanced credit scoring data sets

In 2005, The Basel Committee on Banking Supervision (2005) highlighted the fact that calculations based on historical data made for very safe assets may "not be sufficiently reliable" for estimating the probability of default. The reason for this is that as there are so few defaulted observations, the resulting estimations are likely to be inaccurate. Therefore a need is present for a better understanding of the appropriate modelling techniques for data sets which display a limited number of defaulted observations.

This section has been further sub-divided into problems imbalanced credit scoring data sets pose to modelling (cf. 2.2.1.1.a) and the issue of calibration (cf. 2.2.1.1.b), i.e. how a long-run average that is statistically conservative can be achieved.

2.2.1.1.a PD modelling for imbalanced credit scoring data sets

A wide range of different classification techniques for scoring credit data sets has been proposed in the literature, a non-exhaustive list of which was provided earlier (cf. Chapter 1). In addition, some benchmarking studies have been undertaken to empirically compare the performance of these various techniques (e.g. Baesens et al., 2003), but they

did not focus specifically on how these techniques compare on heavily imbalanced samples, or to what extent any such comparison is affected by the issue of class imbalance. For example, in Baesens et al. (2003), seventeen techniques including both well known techniques such as logistic regression and discriminant analysis and more advanced techniques such as least square support vector machines were compared on eight real-life credit scoring data sets. Although more complicated techniques such as radial basis function least square support vector machines (RBF LS-SVM) and neural networks (NN) yielded good performances in terms of AUC, simpler linear classifiers such as linear discriminant analysis (LDA) and logistic regression (LOG) also gave very good performances. However, there are often conflicting opinions when comparing the conclusions of studies promoting differing techniques. For example, in Yobas et al, (2000), the authors found that linear discriminant analysis (LDA) outperformed neural networks in the prediction of loan default, whereas in Desai et al, (1996), neural networks were reported to actually perform significantly better than LDA. Furthermore, many empirical studies only evaluate a small number of classification techniques on a single credit scoring data set. The data sets used in these empirical studies are also often far smaller and less imbalanced than those data sets used in practice. Hence, the issue of which classification technique to use for credit scoring, particularly with a small number of bad observations, remains a challenging problem (Baesens et al., 2003). In more recent work on the effects of class distribution on the prediction of PD, Crone and Finlay (2011) found that under sampled data sets are inferior to unbalanced and oversampled data sets. However it was also found that the larger the sample size used, the less significant the differences between the methods of balancing were. Their study also incorporated the use of a variety of data mining techniques, including logistic regression, classification and regression trees, linear discriminate analysis and neural networks. From the application of these techniques over a variety of class balances it was found that logistic regression was the least sensitive to balancing. This piece of work is thorough in its empirical design; however it does not assess more novel machine learning techniques in the estimation of default. In the study presented in this thesis, additional techniques such as Gradient Boosting and Random Forests will be adopted to contribute additional value to the literature.

In Yao, (2009) hybrid SVM-based credit scoring models are constructed to evaluate applicant's scoring from an applicant's input features. This paper shows the implications of using machine learning based techniques (SVMs) in a credit scoring context on two widely used credit scoring datasets (Australian credit and German credit) and compares the accuracy of this model against other techniques (LDA, logistic regression and NN). Their findings suggest that the SVM hybrid classifier has the best scoring capability when compared to traditional techniques. Although this is a non-exhaustive study with a bias towards the use of RBF-SVMs it gives a clear basis for the hypothetical use of SVMs in a credit scoring context. The use of the Australian and German credit datasets is also of interest as the same datasets will be utilised in Chapter 4 of this study. A lot can be learned from the empirical setup of this work and will be built upon in this thesis.
In Kennedy, (2011) the suitability of one-class and supervised two-class classification algorithms as a solution to the low-default portfolio problem are evaluated. This study compares a variety of well established credit scoring techniques (e.g. LDA, LOG and k-NN) against the use of a linear kernel SVM. Nine banking datasets are utilised and class imbalance is artificially created by removing 10% of the defaulting observations from the training set after each run. The only issue with this process is that the datasets are comparatively small in size (ranging from 125 - 5,397) which leads this author to believe a process of k-fold cross validation would have been more applicable considering the size of the datasets after a training, validation and test set split are made. However, some merit to this paper are that the findings shown, at least at the 70:30 class split, are comparative to other studies in the area (e.g. Baesens et al. 2003) with no statistical difference in the techniques at this level. As more class imbalance is induced it is shown that logistic regression performs significantly better than Lin-SVM, QDC (Quadratic Bayes Normal) and k-NN. It is also shown that oversampling produces no overall improvement to the best performing two-class classifiers. The findings in this paper lead into the work that will be conducted in this thesis, as several similar techniques and datasets will be employed, alongside the determination of classifier performance on imbalanced data sets.

The topic of which good/bad distribution is the most appropriate in classifying a data set has been discussed in some detail in the machine learning and data mining literature. In Weiss & Provost (2003) it was found that the naturally occurring class distributions in the twenty-five data sets looked at, often did not produce the best-performing classifiers. More specifically, based on the AUC measure (which was preferred over the use of the error rate), it was shown that the optimal class distribution should contain between 50% and 90% minority class examples within the training set. Alternatively, a progressive adaptive sampling strategy for selecting the optimal class distribution is proposed in Provost et al (1999). Whilst this method of class adjustment can be very effective for large data sets, with an adequate number of observations in the minority class of defaulters, in some imbalanced data sets there are only a very small number of loan defaults to begin with.

Various kinds of techniques have been compared in the literature to try and ascertain the most effective way of overcoming a large class imbalance. Chawla et al (2002) proposed a Synthetic Minority Over-sampling technique (SMOTE) which was applied to example data sets in fraud, telecommunications management, and detection of oil spills in satellite images. In Japkowicz (2000) over-sampling and downsizing were compared to the author's own method of "learning by recognition" in order to determine the most effective technique. The findings, however, were inconclusive but demonstrated that both over-sampling the minority class and downsizing the majority class can be very effective. Subsequently Batista (2004) identified ten alternative techniques to deal with class imbalances and trialled them on thirteen data sets. The techniques chosen included a variety of under-sampling and over-sampling methods. Findings suggested that generally oversampling methods provide more accurate results than under-sampling methods. Also, a combination of either SMOTE (Chawla et al, 2002) and Tomek links or SMOTE and ENN (a nearest-neighbour cleaning rule), were proposed.

2.2.1.1.b Imbalanced credit scoring data sets calibration

The purpose of calibration is the assignment of a default probability to each possible score or rating grade values (FMA/OeNB, 2004). With regards to calibration, a

confidence interval-based approach methodology was proposed by Pluto and Tasche (2005) to derive non-zero probabilities of default for credit portfolios with none to very few observed defaults. Their method for estimating imbalanced credit scoring data sets is based on the use of confidence intervals through "the most prudent estimation principle" and incorporating all available quantitative information. Although a variety of confidence levels are discussed the authors suggest that the most intuitively appropriate intervals should be less than 95%. Further to this, a likelihood approach, with a similar methodology to that found in Pluto and Tasche (2005), is applied by Forrest (2005) in the conservative estimation of probabilities of default for imbalanced credit scoring data sets. In this paper, multiple dimensions are used with each dimension representing a different rating grade and each point representing a choice of grade-level PD. A subset of points in this multidimensional space is then identified, conditional on the observed data.

From a regulatory perspective, Benjamin et al (2006) provide a quantitative approach to produce conservative PD estimates when a scarcity of data is present. Centralised PD values are obtained based on the size of the portfolio, the number of observed defaults, and the level of confidence that is placed on the empirical evidence. A comparison can then be made by a financial institution between the values of PD presented (look-up PD) against the weighted average PD of the financial institution's own portfolio. The financial institution then adjusts their PD until their weighted average PD is greater or equal to the presented look-up PDs in the paper.

In Wilde and Jackson (2006), it is shown that probability of default for low-default portfolios can be calculated based on a re-calibration of the CreditRisk+ (cf. Chapter 2.2.1) model to a model of default behaviour similar to that of a Merton model. The challenge of data issues, such as scarcity of defaults, in probability default models is further explored by Dwyer (2006) through the use of Bayesian model validation. A posterior distribution is derived for PD, providing a framework for finding the upper bound for a PD in relation to imbalanced credit scoring data sets. The proposed method allows the determination of when a calibration needs to be recomputed even when a default rate is within the 95% confidence level. Burgt M (2007) looks at the issue of back

testing probability of default rating models in the absence of a sufficient amount of defaults. A method of calibrating these imbalanced credit scoring data sets is presented based on modelling the observed power curve (i.e. Lorenz curve) and deriving the calibration from this curve. This power curve is fitted to a concave function, and the derivative of the concave function and the average default rate are used to perform the calibration.
In Kiefer (2009) the issue of low-default portfolios is looked at with the aim of applying a probability (Bayesian) approach to solving the problem. It is argued that default probability should be represented in a probability distribution in the same way uncertainty is modelled. Hypothetical portfolios of loans with sample sizes ranging from 100-500 are used in the study of the posterior distributions. The results produced in this paper in turn raise issues about how banks should treat estimated default probabilities and how they should be supervised.

In summary, although work has been conducted into the area of imbalanced credit scoring data sets there is still potential for more detailed work to be conducted as gaps still exist e.g. on the modelling level. There is also scope for techniques and methodologies to be used from the Machine Learning literature and applied in a credit scoring context where imbalances in data are present.

2.2.2 Loss Given Default (LGD)

Loss given default (LGD) is the estimated economic loss, expressed as a percentage of exposure, which will be incurred if an obligor goes into default (in other words, 1 – recovery rate in the literature). Producing robust and accurate estimates of potential losses are essential for the efficient allocation of capital within financial organisations for the pricing of credit derivatives and debt instruments (Jankowitsch et al., 2008). Banks are also in the position to gain a competitive advantage if an improvement can be made to their internally made loss-given default forecasts.

Whilst the modelling of probability of default (PD) (cf. Chapter 2.2.1) has been the subject of many studies during the past few decades, literature detailing recovery rates has only emerged more recently. This increase in literature on recovery rates is due to the advent of the new Basel Capital Accord. A detailed review of how credit risk models have developed over the last thirty years on corporate bonds can be found in Altman (2006).

A clear distinction can be made between those models developed for retail credit and corporate credit facilities. As such this section has been sub-divided into four categories distinguishing the literature for retail credit (cf. 2.2.2.a), corporate credit (cf. 2.2.2.b), economic variables (cf. 2.2.2.c) and Downturn LGD (cf. 2.2.2.d).

2.2.2.a LGD models for retail credit

In Bellotti and Crook (2007) alternative regression methods for modelling LGD for credit card loans are evaluated. This work was conducted on a large sample of credit card loans in default and a cross-validation framework using several alternative performance measures are also given. Their findings show that fractional logit regression gives the highest predictive accuracy in terms of mean absolute error (MAE). Another interesting finding is that simple OLS is as good, if not better, than estimating LGD with a Tobit or decision tree approach.

In Somers and Whittaker (2007) quantile regression is applied in two credit risk assessment exercises, including the prediction of LGD for retail mortgages. Their findings suggest that although quantile regression may be usefully applied to solve problems such as forecasting distributions, in estimating LGD however, in terms of R-square the model results are quite poor ranging from 0.05 to a maximum of 0.2.

Grunert J and Weber M (2008) conduct analyses on the distribution of recovery rates and the impact of the quota of collateral, the creditworthiness of the borrower, the size of the company and the intensity of the client relationship on the recovery rate. Their findings show that a high quota of collateral leads to a higher recovery rate.

In Matuszyk et al (2010), a decision tree approach is proposed for modelling the collection process with the use of real data from a UK financial institution. Their findings suggest that a two-stage approach can be used to estimate the class a debtor is in and to estimate the LGD value in each class. A variety of regression models are provided with a weight of evidence (WOE) approach providing the highest coefficient of determination value.

In Hlawatsch and Reichling (2010), two models, a proportional and a marginal decomposition model, for validating relative LGD and absolute losses are developed and presented. Real data from a bank is used in the testing of the models and in-sample and out-of-sample tests are used to test for robustness. Their findings suggest that both their models are applicable without the requirement for first calculating LGD ratings. This is beneficial as LGD ratings are difficult to develop for retail portfolios because of their similar characteristics.

2.2.2.b LGD models for corporate credit

Although few studies have been conducted with the focus on forecasting recoveries, an important study by Moody's KMV gives a dynamic prediction model for LGD modelling called LossCalc (Gupton and Stein, 2005). In this model, over 3000 defaulted loans, bonds and preferred stock observations occurring between the period of 1981 and 2004 are used. The LossCalc model presented is shown to do better than alternative models such as overall historical averages of LGD, and performs well in both out-of-sample and out-of-time predictions. This model allows practitioners to estimate corporate credit losses to a better degree of accuracy than was previously possible.

In the more recent literature on corporate credit, Acharya et al (2007) use an extended set of data on U.S. defaulted firms between 1982 and 1999 to show that creditors of defaulted firms recover significantly lower amounts, in present-value terms, when their particular industry is in distress. They find that not only an economic-downturn effect is present but also a fire-sales effect, also identified by Shleifer and Vsihny (1992). This fire-sales effect means that creditors recover less if the surviving firms are illiquid. The

main finding of this study is that industry conditions at the time of default are robust and economically important determinants of creditor recoveries.

An interesting study by Qi and Zhao (2011) shows the comparison of six statistical approaches to estimation LGD (including regression trees, neural networks and OLS with and without transformations). There findings suggest that non-parametric methods such as neural networks outperform parametric methods such as OLS in terms of model fit and predictive accuracy. It is also shown that the observed values for LGD in the corporate default data set display a bi-modal distribution with focal points around 0 and 1. This paper is limited however by the use of a single corporate defaults data set of a relatively small size (3,751 observations). An extension of this study over multiple data sets and including a variety of additional techniques would therefore add to the validity of the results.

2.2.2.c Economic variables for LGD estimation

It is found in Altman et al. (2005) that when the recovery rates are regressed on the aggregate default rate as an indicator of the aggregate supply of defaulted bonds, a negative relationship is given. However, when macroeconomic variables such as GDP growth, for example, are added as additional explanatory variables, they exhibit low explanatory power for the recovery rates. This indicates that in the prediction of the LGD (recovery rate) at account level, macroeconomic variables do not add anything to the models which only incorporate individual loan-related variables derived from the data.

In Hu and Perraudin (2002), evidence that aggregate quarterly default rates and recovery rates are negatively correlated is presented. This is achieved through the use of Moody's historical bond market data in the period 1971-2000. Their conclusions suggest that recoveries tend to be low when default rates are high. It is also concluded that typical correlations for post 1982 quarters are -22%. Whereas, with respect to the full time period 1971-2000, correlations are typically lower, i.e. -19%.

Caselli et al (2008) verify the existence of a relation between the loss given default rate (LGDR) and macroeconomic variables. Using a sizeable number of bank loans (11,649) concerning the Italian market several models are tested in which LGD is expressed as a linear combination of the explanatory variables. They find that for households, LGDR is more sensitive to the default-to-loan-ratio, the unemployment rate and household consumption. For small to medium enterprises (SMEs) however, LGDR is influenced to a great extent by the GDP growth rate and total number of people employed. The estimation of the model coefficients in this analysis, was achieved by using a simple OLS regression model.

In an extension to their prior work, Bellotti and Crook (2009), add macroeconomic variables to their regression analysis for retail credit cards. The conclusions drawn indicate that although the data used has limitations in terms of the business cycle, adding bank interest rates and unemployment levels as macroeconomic variables into an LGD model yields a better model fit and that these variables are statistically significant explanatory variables.

2.2.2.d Downturn LGD

In terms of estimating Downturn LGD several studies have approached this problem from varying perspectives. For example, in Hartmann-Wendels & Honal (2006) a linear regression model is implemented with the use of a dummy variable to represent Downturn LGD. The findings from this study show that downturn LGD exceeds default-weighted average LGD by eight percent. In Rosch and Scheule (2008), alternative concepts for the calculation of downturn LGD are given on Hong Kong mortgage loan portfolios. Their findings show that the empirical calibration of the downturn LGD agrees with regulatory capital adequacy. Their empirical analysis also highlights that the asset correlations given by the Basel Committee on Banking Supervision (2006) exceed the values empirically estimated and therefore could lead to banks to overprovision for capital.

Further to the papers discussed in this section, the following additional papers provide information on other areas of loss given default modelling; Benzschawel et al. (2011);

Jacobs and Karagozoglu (2011); Sigrist and Stahel (2010); Luo and Shevchenko (2010); Bastos (2010); Hlawatsch and Ostrowski (2010); Li (2010); Chalupka and Kopecsni (2009).

2.2.3 Exposure at Default (EAD)

In this section, a literature review of the current work conducted in the area of EAD is given. To date the main focus of the literature has been conducted on corporate lending as opposed to retail lending (i.e. consumer credit, e.g. through credit cards), with only more recent studies taking into account the implications for retail lending. We will begin by identifying these corporate lending studies and go on to look at the current retail lending literature. Note that, in this thesis, the term Loan Equivalency Factor (LEQ) is used interchangeably with the term credit conversion factor (CCF) as CCF is referred to as LEQ in U.S. publications.

2.2.3.a EAD models for corporate credit

To our knowledge, the earliest reported work on EAD modelling for corporate lending was on data from Chase Manhattan Bank in 1994 (Araten and Keisman, 1994, via Jacobs, 2008), which was later updated in 2003. In this study, 104 revolving credit facilities were analysed and LEQs were directly estimated for facilities with rating grades of BB or below. The conclusions drawn from this study identify a correlation where the greater the risk rating and tenor, the larger the LEQ factor would be. In a similar vein, Asarnow and Marker (1995) looked at utilisation patterns for revolving commitments at Citibank over a 5 year period (1988 - 1993). In this study, the importance of credit ratings in the estimation of the LEQs, in particular commitments with speculative ratings, is shown.

More recently, Araten and Jacobs (2001) used six years of data between 1994 and 2000 from Chase bank to calculate values for the LEQ factor. It was found that the estimated LEQs calculated increased the closer the period of time to default and with better risk

categories. It was also found that the distribution of the LEQ value had significant concentrations around the 0 and 1 values, giving a two-peaked distribution.
In the most recent work on corporate lending, Cespedes e*t al* (2010) look at the issue of modelling wrong way risk in the estimation of an alpha multiplier (the definition of a portfolio's alpha is: total economic capital divided by the economic capital when counterparty exposures are equal to expected positive exposure (EPE)). The alpha value typically ranges from 1.1 for large global portfolios to over 2.5 for new users with concentrated exposures. Wrong way risk is defined here as the correlation between exposures and defaults in a given credit portfolio. Their paper gives a computationally efficient and robust approach to the modelling of the alpha multiplier and stress-testing wrong way risk. This is achieved through leveraging underlying counterparty potential future exposure (PFE) simulations that are also used for credit limits and risk management. An application of the methodology is provided on a realistic bank trading portfolio with the results indicating that the alpha remains at or below 1.2 for conservative correlation assumptions. Prior to this, Sufi (2009) looked at the use of credit lines to corporations. The conclusions drawn show that the flexibility given to firms by the use of unfunded commitments leads to a moral hazard problem. To tackle this, banks tend to impose strict agreements, and only lend to borrowers with historically high profitability.

2.2.3.b EAD models for retail credit

With regards to retail lending, Taplin et al (2007) focus on the treatment of exposure at default (EAD) for undrawn lines of credit through the use of data from defaulted credit cards at BankWest. Although the main focus of this article is in the context of retail lines of credit the concepts developed and points made can be generalised and applied to the treatment of all lines of credit. The findings show that in the context of EAD modelling the prescribed formulation of CCF in Basel II can be inappropriate whereas modelling EAD directly is more statistically logical. There conclusions also indicate that a more appropriate single parameter model for EAD is $EAD = \beta L$ where L is defined as the limit amount and β is defined as an 'uplift factor'. The use of β estimates the amount EAD is expected to exceed L, and can be varied dependent on account characteristics.

They do however warn that in general a single parameter model, be it the use of a CCF or an uplift factor is too simplistic. Strong arguments are given throughout this study indicating that practitioners should take care and apply common sense to their models for estimating EAD and take advantage of the flexibility offered by the Basel II Accord.

In Qi (2009) borrower and account information for unsecured credit card defaults from a large US lender are used to calculate and model CCF (referred to as LEQ in the US). The findings suggest that borrowers' attributes such as credit score, aggregate bankcard balance, aggregate bankcard credit line utilization rate, number of recent credit inquiries, and number of open retail accounts are significant drivers of CCF for accounts current one year prior to default. It is also found that borrowers are more likely to draw down additional funds as they approach default.
In Valvonis (2008) the issues related to the estimation of EAD and CCF are discussed in detail as well as the EAD risk drivers (EADRDs). The findings suggest that many issues pertaining to EAD modelling remain unanswered such as the issue of the stringent supervisory requirements banks are under in their calculations of EAD. It is also shown that point densities for the majority of realised CCFs occur around 0 and 1, and it is suggested that logit or probit regression models could indeed be appropriate here.

In the academic and regulatory literature, on the other hand, there has been little work done on the estimation of EAD and the appropriate models required. The majority of work to date has been done on modelling exposure at default (EAD) for defaulted firms. Most notably, in Jacobs (2008), a variety of explanatory variables are investigated with various measures of EAD risk derived and compared. Also, a multiple regression model in the generalised linear class is built. The findings suggest that there is weak evidence for counter-cyclicality in EAD and utilization has a strong inverse relationship to EAD. As with Asarnow and Marker (1995), the risk rating is found to have some explanatory power in the estimation of the EAD. Similarly to Jacobs (2008), other academic work that has been conducted in this area has also focused on corporate lending as opposed to retail lending. In Jiménez *et al* (2009), LEQ factors for revolving commitments in the Spanish credit register are studied over the period of the last two decades for corporate lending.

The conclusions drawn from this work are that the firms that go into default have much higher credit line usage rates and EAD values up to five years prior to their default than non-defaulting facilities. Variations in EAD are also identified due to collateralisation, maturity and credit line size.

In the regulatory paper by Moral (2006), a variety of suggested regulatory guidelines relating to the CCF and its estimation are given. Several approaches are detailed for the calculation of the CCF based on the period of time used as the reference date. The potential approaches given for the selection of the time period are the fixed and variable time horizon approaches and the cohort approach. The guidelines given also identify a selection of quantitative risk drivers for the estimation of the EAD, which include the commitment size of the exposure, the drawn and undrawn amounts, the credit percentage, time to default, rating class and status of the facility. The risk drivers suggested in Moral (2006) will be analysed and further built upon in our study. A report by Gruber and Parchert (2006) also discusses the estimation of the EAD for both balance-sheet and off-balance sheet financial products. Several internal approaches for the estimation of EAD for derivative products are identified. These techniques are the variance-covariance approach for calculating the potential future exposure (PFE) and Monte-Carlo one and multi-step approaches. It is proposed that to avoid the shortcomings of the regulatory methods, more elaborate techniques such as Monte-Carlo techniques can be applied for estimating the EAD for derivative products.

In summary, although more recent studies on EAD modelling have become available for retail lending, there is a clear need to further develop our understanding of the risk drivers and appropriate EAD modelling techniques for consumer credit. Hence, in this paper, we will investigate both using a real-world data set of credit card defaults.

2.3 Summary of Literature Review

In summary, as shown in this chapter a wide range of modelling work has been conducted in the field of credit risk modelling, with particular attention paid to that of probability of default (PD) modelling. Since the advent of the Basel II capital Accord however there has become an even greater need for the development of suitable and robust estimation techniques for loss given default (LGD) and exposure at default (EAD), as well as a more comprehensive review of the appropriate techniques to use when a scarcity of defaults is present (imbalanced data sets). It is therefore the focus of this thesis to provide a better understanding of the classification and regression techniques required for the prediction of imbalanced credit scoring data sets, LGD and EAD as well as providing robust statistical results.

In the next chapter, a detailed explanation of each of the techniques applied in this thesis will be presented.

Chapter 3

3 Classification and Regression Techniques

This thesis analyses a variety of established and novel classification and regression techniques in the estimation of the three components of the minimum capital requirement, PD, LGD and EAD.

Classification is defined as the process of assigning a given piece of input data into one of a given number of categories. In terms of Probability of Default (PD) modelling, classification techniques are applied as the purpose of PD modelling is to estimate the likelihood that a loan will not be repaid and will fall into default, this requires the classification of loan applicants into two classes, i.e. good payers (i.e., those who are likely to keep up with their repayments) and bad payers (i.e., those who are likely to default on their loans). Regression analysis estimates the conditional expectation of a dependent variable given a linear or non-linear combination of a set of independent variables. This is therefore appropriate for use in the estimation of Loss Given Default (LGD) and Exposure at Default (EAD) where the goal is to determine their conditional expectations given a set of independent variables.
The literature review section of this thesis (cf. Chapter 2.1) identified current and potentially applicable classification and regression techniques to the field of credit risk modelling. Therefore, this thesis aims to apply the most prevalent techniques identified with the aim of finding the most appropriate techniques in the estimation of PD, LGD and EAD.

In this chapter a brief explanation of each of the techniques applied in this thesis is presented with citations given to their full derivation. (N.B. some of the techniques described have applications in both classification and regression analysis. Where this is the case the technique is only described in the classification section.)

3.1 Overview of Classification Techniques

In Chapter 4, of this thesis we aim to compare the performance of a wide range of classification techniques within a credit scoring context, thereby assessing to what extent they are affected by increasing class imbalance. For the purpose of this thesis, ten classifiers have been selected which provide a balance between well-established credit scoring techniques such as logistic regression, decision trees and neural networks, and newly developed machine learning techniques such as least square support vector machines, gradient boosting and random forests. A brief explanation of each of the classification techniques applied in this thesis is presented below. This section details the basic concepts and functioning of a selection of well used classification methods. Although other classification techniques have been identified in the literature prior (c.f. Chapter 2) we believe the selection made here discusses the most prevalent and pertinent to the research topics presented in this thesis.

3.1.1 Logistic Regression (LOGIT & CLOGIT)

In Chapter 4 of this thesis, we will be focusing on the binary response of whether a creditor turns out to be a good or bad payer (i.e. non-defaulter vs. defaulter). For this binary response model, the response variable, y, can take on one of two possible values; i.e., $y = 0$ if the customer is a bad payer, $y = 1$ if he/she is a good payer. Let us assume $\mathbf{x}$ is a column vector of M explanatory variables and $\pi = \mathrm{P}(y = 1 | \mathbf{x})$ is the response probability to be modelled. The logistic regression model then takes the form:

$$\mathrm{logit}(\pi) \equiv \log\left(\frac{\pi}{1-\pi}\right) = \alpha + \beta^T \mathbf{x} \qquad (3.1)$$

where α is the intercept parameter and β^T contains the variable coefficients (Hosmer and Stanley, 2000).

The cumulative logit model (see e.g. Walker and Duncan, 1967) is simply an extension of the binary two-class logit model which allows for an ordered discrete outcome with more than 2 levels $(k > 2)$:

$$P(\text{class} \leq j) = \frac{1}{1+e^{-\left(d_j + b_1 x_1 + b_2 x_2 + \ldots + b_n x_n\right)}}, \tag{3.2}$$

$$j = 1, 2, \ldots, k-1.$$

The cumulative probability, denoted by $P(\text{class} \leq j)$, refers to the sum of the probabilities for the occurrence of response levels up to and including the j th level of y. The main advantage of logistic regression is the fact that it is a non-parametric classification technique as no prior assumptions are made with regard to the probability distribution of the given attributes.

3.1.2 Linear and Quadratic Discriminant Analysis (LDA & QDA)

Discriminant analysis assigns an observation to the response, y ($y \in \{0,1\}$), with the largest posterior probability; i.e., classify into class 0 if $p(0|\mathbf{x}) > p(1|\mathbf{x})$, or class 1 if the reverse is true. According to Bayes' theorem, these posterior probabilities are given by:

$$p(y|\mathbf{x}) = \frac{p(\mathbf{x}|y)p(y)}{p(\mathbf{x})}. \tag{3.3}$$

Assuming now that the class-conditional distributions $p(\mathbf{x}|y=0)$, $p(\mathbf{x}|y=1)$ are multivariate normal distributions with mean vector $\boldsymbol{\mu}_0$, $\boldsymbol{\mu}_1$, and covariance matrix $\boldsymbol{\Sigma}_0$, $\boldsymbol{\Sigma}_1$, respectively, the classification rule becomes: classify as $y=0$ if the following is satisfied:

$$\begin{aligned}&(\mathbf{x}-\boldsymbol{\mu}_0)^T \sum_0^{-1}(\mathbf{x}-\boldsymbol{\mu}_0) \; - \; (\mathbf{x}-\boldsymbol{\mu}_1)^T \sum_1^{-1}(\mathbf{x}-\boldsymbol{\mu}_1) \\ &< 2\Big(\log\big(P(y=0)-\log\big(P(y=1)\big)\big)\Big)+\log|\boldsymbol{\Sigma}_1|-\log|\boldsymbol{\Sigma}_0|\end{aligned}. \tag{3.4}$$

Linear discriminant analysis is then obtained if the simplifying assumption is made that both covariance matrices are equal, i.e. $\Sigma_0 = \Sigma_1 = \Sigma$, which has the effect of cancelling out the quadratic terms in the expression above.

3.1.3 Neural Networks (NN)

Neural networks (NN) are mathematical representations modelled on the functionality of the human brain (Bishop, 1995). The added benefit of a NN is its flexibility in modelling virtually any non-linear association between input variables and target variable. Although various architectures have been proposed, this thesis focuses on probably the most widely used type of NN, i.e. the Multilayer Perceptron (MLP). A MLP is typically composed of an input layer (consisting of neurons for all input variables), a hidden layer (consisting of any number of hidden neurons), and an output layer (in our case, one neuron). Each neuron processes its inputs and transmits its output value to the neurons in the subsequent layer. Each such connection between neurons is assigned a weight during training. The output of hidden neuron i is computed by applying an activation function $f^{(1)}$ (for example the logistic function) to the weighted inputs and its bias term $b_i^{(1)}$:

$$h_i = f^{(1)}\left(b_i^{(1)} + \sum_{j=1}^{n} \mathbf{W}_{ij} x_j\right), \tag{3.5}$$

where $\mathbf{W}$ represents a weight matrix in which $\mathbf{W}_{ij}$ denotes the weight connecting input j to hidden neuron i. For the analysis conducted in this thesis, a binary prediction will be made; hence, for the activation function in the output layer, we will be using the logistic (sigmoid) activation function, $f^{(2)}(x) = \dfrac{1}{1+e^{-x}}$ to obtain a response probability:

$$\pi = f^{(2)}\left(b^{(2)} + \sum_{j=1}^{n_h} \mathbf{v}_j h_j\right), \tag{3.6}$$

with n_h the number of hidden neurons and $\mathbf{v}$ the weight vector where $\mathbf{v}_j$ represents the weight connecting hidden neuron j to the output neuron. Examples of other transfer functions that are commonly used are the hyperbolic tangent $f(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$ and the linear transfer function $f(x) = x$.

During model estimation, the weights of the network are first randomly initialised and then iteratively adjusted so as to minimise an objective function, e.g. the sum of squared errors (possibly accompanied by a regularisation term to prevent over-fitting). This iterative procedure can be based on simple gradient descent learning or more sophisticated optimisation methods such as Levenberg-Marquardt or Quasi-Newton. The number of hidden neurons can be determined through a grid search based on validation set performance.

3.1.4 Least Square Support Vector Machines (LS-SVM)

Support vector machines (SVMs) are a set of powerful supervised learning techniques used for classification and regression. Their basic principle, when applied as a classifier, is to construct a maximum-margin separating hyperplane in some transformed feature space. Rather than requiring one to specify the exact transformation though, they use the principle of kernel substitution to turn them into a general (non-linear) model. The least square support vector machine (LS-SVM) proposed by Suykens, *et al* (2002) is a further adaptation of Vapnik's original SVM formulation which leads to solving linear KKT (Karush-Kuhn-Tucker) systems (rather than a more complex quadratic programming problem). The optimisation problem for the LS-SVM is defined as:

$$\min_{\mathbf{w},b,\mathbf{e}} J(\mathbf{w},b,\mathbf{e}) = \frac{1}{2}\mathbf{w}^{\mathbf{T}}\mathbf{w} + \gamma\frac{1}{2}\sum_{i=1}^{N} e_i^2\,, \tag{3.7}$$

subject to the following equality constraints:

$$y_i\left[\mathbf{w}^{\mathbf{T}}\varphi(\mathbf{x}_i)+b\right]=1-e_i,\ \ i=1,\ldots,l. \tag{3.8}$$

Where $\mathbf{w}$ is the weight vector in primal space, γ is the regularisation parameter, and $y_i=+1$ or -1 for good (bad) payers, respectively (Suykens, *et al* 2002). A solution can then be obtained after constructing the Lagrangian, and choosing a particular kernel function $K\left(\mathbf{x},\mathbf{x}_i\right)$ that computes inner products in the transformed space, based on which a classifier of the following form is obtained:

$$y=\operatorname{sign}\left[\sum_{i=1}^{l}\alpha_i y_i K\left(\mathbf{x},\mathbf{x}_i\right)+b\right], \tag{3.9}$$

whereby $K\left(\mathbf{x},\mathbf{x}_i\right)=\varphi\left(\mathbf{x}\right)^T\varphi\left(\mathbf{x}_i\right)$ is taken to be a positive definite kernel satisfying the Mercer theorem. The hyper parameter γ for the LS-SVM classification technique could, for example, be tuned using 10-fold cross validation.

3.1.5 Decision Trees (C4.5)

Classification and regression trees are decision tree models, for a categorical or continuous dependent variable, respectively, that recursively partition the original learning sample into smaller subsamples, so that some impurity criterion $i()$ for the resulting node segments is reduced (Breiman, *et al* (1984). To grow the tree, one typically uses a greedy algorithm that, at each node t, evaluates a large set of candidate variable splits so as to find the 'best' split, i.e. the split s that maximises the weighted decrease in impurity:

$$\Delta i\left(s,t\right)=i\left(t\right)-p_L i\left(t_L\right)-p_R i\left(t_R\right). \tag{3.10}$$

Where p_L and p_R denote the proportions of observations associated with node t that are sent to the left child node t_L or right child node t_R, respectively. A decision tree consists of internal nodes that specify tests on individual input variables or attributes that split the data into smaller subsets, and a series of leaf nodes assigning a class to each of the observations in the resulting segments. For Chapter 4, we chose the popular decision tree classifier C4.5, which builds decision trees using the concept of information entropy (Quinlan, 1993). The entropy of a sample *S* of classified observations is given by:

$$Entropy(S) = -p_1 \log_2(p_1) - p_0 \log_2(p_0), \tag{3.11}$$

where p_1 and p_0 are the proportions of the class values 1 and 0 in the sample *S*, respectively. C4.5 examines the normalised information gain (entropy difference) that results from choosing an attribute for splitting the data. The attribute with the highest normalised information gain is the one used to make the decision. The algorithm then recurs on the smaller subsets.

3.1.6 Memory Based Reasoning (k-NN)

The k-nearest neighbours algorithm (k-NN) classifies a data point by taking a majority vote of its *k* most similar data points (Hastie, *et al* 2001). The similarity measure used in this thesis is the Euclidean distance between the two points:

$$d(\mathbf{x}_i, \mathbf{x}_j) = \left\| \mathbf{x}_i - \mathbf{x}_j \right\| = \left[\left(\mathbf{x}_i - \mathbf{x}_j \right)^T \left(\mathbf{x}_i - \mathbf{x}_j \right) \right]^{1/2}. \tag{3.12}$$

One of the major disadvantages of the k-nearest neighbour classifier is the large requirement on computing power as for classifying an object, the distance between it and all the objects in the training set has to be calculated. Furthermore, when many irrelevant attributes are present, the classification performance may degrade when observations have distant values for these attributes (Baesens B, 2003a).

3.1.7 Random Forests

Random forests are defined as a group of un-pruned classification or regression trees, trained on bootstrap samples of the training data using random feature selection in the process of tree generation. After a large number of trees have been generated, each tree votes for the most popular class. These tree voting procedures are collectively defined as random forests. A more detailed explanation of how to train a random forest can be found in Breiman (2001). For the Random Forests classification technique two parameters require tuning. These are the number of trees and the number of attributes used to grow each tree.

The two meta-parameters that can be set for the Random Forests classification technique are: the number of trees in the forest and the number of attributes (features) used to grow each tree. In the typical construction of a tree, the training set is randomly sampled, then a random number of attributes is chosen with the attribute with the most information gain comprising each node. The tree is then grown until no more nodes can be created due to information loss

3.1.8 Gradient Boosting

Gradient boosting (Friedman, 2001, 2002) is an ensemble algorithm that improves the accuracy of a predictive function through incremental minimisation of the error term. After the initial base learner (most commonly a tree) is grown, each tree in the series is fit to the so-called "pseudo residuals" of the prediction from the earlier trees with the purpose of reducing the error. The estimated probabilities are adjusted by weight estimates, and the weight estimates are increased when the previous model misclassified a response. This leads to the following model:

$$F(\mathbf{x}) = G_0 + \beta_1 T_1(\mathbf{x}) + \beta_2 T_2(\mathbf{x}) + \ldots + \beta_u T_u(\mathbf{x}), \tag{3.13}$$

where G_0 equals the first value for the series, $T_1,\ldots,T_u$ are the trees fitted to the pseudo-residuals, and β_i are coefficients for the respective tree nodes computed by the Gradient Boosting algorithm. A more detailed explanation of gradient boosting can be found in Friedman (2001) and Friedman (2002). The meta-parameters which require tuning for a Gradient Boosting classifier are the number of iterations and the maximum branch used in the splitting rule. The number of iterations specifies the number of terms in the boosting series, for a binary target the number of iterations determines the number of trees. The maximum branch parameter determines the maximum number of branches that the splitting rule produces from one node, a suitable number for this parameter is 2, a binary split.

3.2 Overview of Regression Techniques

Whereas in the previous section we looked at the proposed classification techniques for PD modelling, in this section we will detail the proposed regression techniques to be implemented in the modelling of LGD and EAD. The experiments comprise a selection of one-stage and two-stage techniques. One-stage techniques can be divided into linear and non-linear techniques. The linear techniques included in Chapter 5 and 6, model the (original or transformed) dependent variable as a linear function of the independent variables whereas non-linear techniques fit a non-linear model to the data set. Two-stage models are a combination of the aforementioned one-stage models. These either combine the comprehensibility of an OLS model with the added predictive power of a non-linear technique, or they use one model to first discriminate between zero- and higher LGDs and a second model to estimate LGD for the subpopulation of non-zero LGDs.

A regression technique fits a model $y = f(\mathbf{x}) + e$ onto a data set, where y is the dependent variable, $\mathbf{x}$ are the independent variables and e is the residual.

The following Table (TABLE 3.1) details the regression techniques used in Chapter 5 for LGD estimation and Chapter 6 for EAD estimation.

Regression Techniques
LGD
Linear
Ordinary Least Squares (OLS)
Ordinary Least Squares with Beta Transformation (B-OLS)
Beta Regression (BR)
Ordinary Least Squares with Box-Cox Transformation (BC-OLS)
Non-linear
Regression Trees (RT)
Least Square Support Vector Machines (LS-SVM)
Neural Networks (NN)

Log+(non-)linear
Logistic regression + OLS, B-OLS, BR, BC-OLS, RT, LS-SVM or NN
Linear+non-linear
Ordinary Least Squares + Regression Trees (OLS+RT)
Ordinary Least Squares + Least Square Support Vector Machines (OLS+LSSVM)
Ordinary Least Squares + Neural Networks (OLS+NN)
EAD
Ordinary Least Squares (OLS)
Ordinary Least Squares with Beta Transformation (B-OLS)
Binary Logistic Regression (LOGIT)
Cumulative Logistic Regression (CLOGIT)

TABLE 3.1: Regression techniques used for LGD and EAD modelling

3.2.1 Ordinary Least Squares (OLS)

Ordinary least squares regression (Draper & Smith, 1998) is the most common technique to find optimal parameters $\mathbf{b}^T = [b_0, b_1, b_2, \ldots, b_n]$ to fit a linear model to a data set:

$$y = \mathbf{b}^T \mathbf{x}, \tag{3.14}$$

where $\mathbf{x}^T = [1, x_1, x_2, \ldots, x_n]$. OLS approaches this problem by minimising the sum of squared residuals:

$$\sum_{i=1}^{l} (e_i)^2 = \sum_{i=1}^{l} \left(y_i - \mathbf{b}^T \mathbf{x}_i \right)^2 . \tag{3.15}$$

By taking the derivative of this expression and subsequently setting the derivative equal to zero:

$$\sum_{i=1}^{l} \left(y_i - \mathbf{b}^T \mathbf{x}_i \right) \mathbf{x}_i^T = 0, \tag{3.16}$$

the model parameters $\mathbf{b}$ can be retrieved as:

$$\mathbf{b} = \left(\mathbf{X}^T\mathbf{X}\right)^{-1}\mathbf{X}^T\mathbf{y}, \tag{3.17}$$

with $\mathbf{X}^T = \left[\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_l\right]$ and $\mathbf{y} = \left[y_1, y_2, \ldots, y_l\right]^T$.

3.2.2 Ordinary Least Squares with Beta transformation (B-OLS)

Whereas OLS regression tests generally assume normality of the dependent variable y, the empirical distribution of LGD can often be approximated more accurately by a Beta distribution (Gupton & Stein, 2002). Assuming that y is constrained to the open interval $(0,1)$, the cumulative distribution function (CDF) of a Beta distribution is given by:

$$\beta(y; a, b) = \frac{\Gamma(a+b)}{\Gamma(a)\Gamma(b)}\int_0^y v^{a-1}(1-v)^{b-1}\,dv, \tag{3.18}$$

where $\Gamma()$ denotes the well-known Gamma function, and a and b are two shape parameters, which can be estimated from the sample mean μ and variance σ^2 using the method of the moments, i.e.:

$$a = \frac{\mu^2(1-\mu)}{\sigma^2} - \mu\,;\; b = a\left(\frac{1}{\mu} - 1\right). \tag{3.19}$$

A potential solution to improve model fit therefore is to estimate an OLS model for a transformed dependent variable $y_i^* = N^{-1}\left(\beta(y_i; a, b)\right)$ $(i = 1, \ldots, l)$, in which $N^{-1}()$ denotes the inverse of the standard normal CDF. The predictions by the OLS model are then transformed back through the standard normal CDF and the inverse of the fitted Beta CDF to get the actual LGD estimates.

3.2.3 Beta Regression (BR)

Instead of performing a Beta transformation prior to fitting an OLS model, an alternative Beta regression approach is outlined in Smithson & Verkuilen, (2006). Their preferred model for estimating a dependent variable bounded between 0 and 1 is closely related to the class of generalised linear models and allows for a dependent variable that is Beta-distributed conditional on the covariates. Instead of the usual parameterisation though of the Beta distribution, with shape parameters a and b, they propose an alternative parameterisation involving a location parameter μ and a precision parameter ϕ, by letting:

$$\mu = \frac{a}{a+b};\ \phi = a+b. \tag{3.20}$$

It can be easily shown that the first parameter is indeed the mean of a $\beta(a,b)$-distributed variable, whereas $\sigma^2 = \frac{\mu(1-\mu)}{(\phi+1)}$, so for fixed μ, the variance (dispersion) increases with smaller ϕ.

Two link functions mapping the unbounded input space of the linear predictor into the required value range for both parameters are then chosen, viz. the logit link function for the location parameter (as its value must be squeezed into the open unit interval) and a log function for the precision parameter (which must be strictly positive), resulting in the following sub models:

$$\mu_i = E(y_i \mid \mathbf{x}_i) = \frac{e^{\mathbf{b}^T \mathbf{x}_i}}{1+e^{\mathbf{b}^T \mathbf{x}_i}}. \tag{3.21}$$

$$\phi_i = e^{-\mathbf{d}^T \mathbf{x}_i}$$

This particular parameterisation offers the advantage of producing more intuitive variable coefficients (as the two rows of coefficients, $\mathbf{b}^T$ and $\mathbf{d}^T$, provide an indication of the effect on the estimate itself and its precision, respectively). By further selecting which

variables to include in (or exclude from) the second sub model, one can explicitly model heteroskedasticity. The resulting log-likelihood function is then used to compute maximum-likelihood estimators for all model parameters.

3.2.4 Ordinary Least Squares with Box-Cox transformation (BC-OLS)

The aim of the family of Box-Cox transformations (Box & Cox, 1964) is to make the residuals of the regression model more homoskedastic and closer to a normal distribution. The Box-Cox transformation on the dependent variable y_i takes the form

$$\begin{cases} \dfrac{\left(\left(y_i + c\right)^{\lambda} - 1\right)}{\lambda} & \text{if } \lambda \neq 0 \\ \log\left(y_i + c\right) & \text{if } \lambda = 0 \end{cases}, \tag{3.22}$$

with power parameter λ and parameter c. If needed, the value of c can be set to a non-zero value to rescale y so that it becomes strictly positive. After a model is built on the transformed dependent variable using OLS, the predicted values can be transformed back to their original value range.

3.2.5 Regression trees (RT)

In Section 3.1.5 we looked at the application of decision trees for classification problems. Decision trees can also be used for regression analysis where they are designed to approximate real-valued functions as apposed to a classification task. A commonly applied impurity measure $i(t)$ for regression trees is the mean squared error or variance for the subset of observations falling into node t. Alternatively, a split may be chosen based on the p-value of an ANOVA F-test comparing between-sample variances against within-sample variances for the subsamples associated with its respective child nodes (ProbF criterion).

3.2.6 Artificial Neural Networks (ANN)

Section 3.1.3 details the implementation of Neural Networks for classification problems. In terms of regression, Neural Networks produce an output value by feeding inputs through a network whose subsequent nodes apply some chosen activation function to a weighted sum of incoming values. The type of NN used in Chapter 5 of this thesis is the popular multilayer perceptron (MLP).

3.2.7 Least Square Support Vector Machines (LS-SVM)

Section 3.1.4 details the implementation of Least Square Support Vector Machines for classification problems. In terms of regression, Least Square Support Vector Machines implicitly map the input space to a kernel-induced high-dimensional feature space in which a linear relationship is fitted.

3.2.8 Linear regression + non-linear regression

Techniques such as Neural Networks and Support Vector Machines are often seen as "black box" techniques meaning that the model obtained is not understandable in terms of physical parameters. This is an obvious issue when applying these techniques to a credit risk modelling scenario where physical parameters are required. To solve this issue we propose the use of a two-stage approach to combine the good comprehensibility of OLS with the predictive power of a non-linear regression technique (Van Gestel, *et al* 2005). In the first stage, an ordinary least squares regression model is built:

$$y = \mathbf{b}^T \mathbf{x} + e\,. \tag{3.23}$$

In the second stage, the residuals e of this linear model:

$$e = f\left(\mathbf{x}\right) + e^*, \tag{3.24}$$

are estimated with a non-linear regression model $f(\mathbf{x})$ in order to further improve the predictive ability of the model. Doing so, the model takes the following form:

$$y = \mathbf{b}^T\mathbf{x} + f(\mathbf{x}) + e^*. \tag{3.25}$$

Where e^* are the new residuals of estimating e. A combination of OLS with RT, LSSVM and NN is assessed in this thesis.

3.2.9 Logistic regression + (non)linear regression

The LGD distribution is often characterised by a large peak around $LGD = 0$. This non-normal distribution can lead to inaccurate regression models. This proposed two-stage technique attempts to resolve this issue by modelling the peak separately from the rest. Therefore, the first stage of this two-stage model consists of a logistic regression to estimate whether $LGD \leq 0$ or $LGD > 0$.
In a second stage the mean of the observed values of the peak is used as prediction in the first case and a one-stage (non)linear regression model is used to provide a prediction in the second case. The latter is trained on part of the data set, i.e. those observations that have an $LGD > 0$. More specifically, a logistic regression results in an estimate of the probability P of being in the peak:

$$P = \frac{1}{1 + e^{-(\mathbf{b}^T\mathbf{x})}}, \tag{3.26}$$

with $(1-P)$ as the probability of not being in the peak. An estimate for LGD is then obtained by:

$$y = P.\bar{y}_{peak} + (1-P).f(\mathbf{x}) + e, \tag{3.27}$$

where $\bar{y}_{peak}$ is the mean of the values of $y \leq 0$, which is equitable to 0, and $f(\mathbf{x})$ is a one-stage (non)linear regression model, built on those observations only that are not in the peak. A combination of logistic regression with all aforementioned one-stage techniques as described above, is assessed is this thesis (Matuszyk et al 2010).

Chapter 4

4 Building default prediction models for imbalanced credit scoring data sets

In this chapter, we set out to compare several techniques that can be used in the analysis of imbalanced credit scoring data sets. In a credit scoring context, imbalanced data sets occur as the number of defaulting loans in a data set is usually much lower than the number of observations that do not default.

However, some techniques may not be able to adequately cope with these imbalanced data sets therefore, the objective is to compare a variety of techniques performances' over differing sizes of class distribution. As well as evaluating traditional classification techniques such as logistic regression, neural networks and decision trees, this chapter will also explore the suitability of gradient boosting, least square support vector machines and random forests for loan default prediction. These particular techniques have been selected due to either their proven ability within the credit scoring domain (c.f. TABLE1.1) or their similar applications in other fields which can be transferred to a credit scoring context (c.f. Literature Review). The purpose of this study is to compare widely used credit scoring techniques against novel machine learning techniques to identify whether any improvement can be made over traditional techniques when a class imbalance is present.

Five real-world credit scoring data sets have been adapted to mimic imbalanced data sets and are used to build classifiers and test their performance. In our experiments, we

progressively increase class imbalance in each of these data sets by randomly under-sampling the minority class of defaulters, so as to identify to what extent the predictive power of the respective techniques is adversely affected.
The performance criterion chosen to measure this effect is the area under the receiver operating characteristic curve (AUC); Friedman's statistic and Nemenyi post-hoc tests are used to test for significance of AUC differences between techniques.
The results from this empirical study indicate that the Random Forest and Gradient Boosting classifiers perform very well in a credit scoring context and are able to cope comparatively well with pronounced class imbalances in these data sets. We also find that, when faced with a large class imbalance, the support vector machines and quadratic discriminant analysis perform significantly worse than the best performing classifiers.

The remainder of this chapter is organised as follows. Section 4.2 gives a list overview of the examined classification techniques (a more detailed explanation of each of the techniques used in this chapter can be found in Chapter 3). Section 4.3 details the empirical set up, data sets used and the criteria used for comparing the classification performance. Section 4.4 the results of our experiments are presented and discussed. Finally section 4.5 gives the conclusions that can be drawn from the study and recommendations for further research work will be outlined.

4.1 Introduction

A detailed background and introduction to the topic of estimating Probability of Default (PD) for imbalanced credit scoring data sets along with motivations for the work can be found in Chapter 1 of this thesis.

.

4.2 Overview of classification techniques

This chapter aims to compare the performance of a wide range of classification techniques within a credit scoring context, thereby assessing to what extent they are affected by increasing class imbalance. For the purpose of this chapter, ten classifiers have been selected which provide a comparison between well-established credit scoring techniques such as logistic regression, decision trees and neural networks, and newly developed machine learning techniques such as least square support vector machines, gradient boosting and random forests. An explanation of each of the techniques applied in this chapter can be found in Chapter 3.

The techniques used in this chapter are as follows:

Classification Technique
1. Logistic Regression
2. Linear Discriminant Analysis
3. Quadratic Discriminant Analysis
4. Neural Networks (Multi-layer Perceptron)
5. Least Square Support Vector Machines (LS-SVMs)
6. C4.5 – Decision Trees
7. k-NN10 (Memory Based Reasoning)
8. k-NN100 (Memory Based Reasoning)
9. Random Forests
10. Gradient Boosting

TABLE 4.1: List of classification techniques

4.3 Experimental set-up and data sets

4.3.1 Data set characteristics

The characteristics of the data sets used in evaluating the performance of the aforementioned classification techniques are given below in TABLE 4.2. (The independent variables available in each data set are presented in APPENDIX A1 at the end of this thesis). The Bene1 and Bene2 data sets were obtained from two major financial institutions in the Benelux region. For these two data sets, a bad customer was defined as someone who had missed three consecutive months of payments. The German credit data set and the Australian Credit data set are publicly available at the UCI repository (http://kdd.ics.uci.edu/). The Behav data set was also acquired from a Benelux institution. As the data sets used vary in size, from 547 to 7,190, and the data sets will be further reduced, with the under sampling of the bad observations to create larger class imbalances, a process of 10-fold cross validation will be applied on the full data set.

	Inputs	Data set size	Goods/Bads
Bene1	27	2974	70/30*
Bene2	27	7190	70/30
Austr	14	547	70/30*
Behav	60	1197	70/30*
Germ	20	1000	70/30

TABLE 4.2: Characteristics of credit scoring data sets

* Altered data set class distribution, Bene1 original distribution was 66.6% good observations, 33.3% bad observations, Austr original distribution was 55.5% good observations, 44.5% bad observations and the Behav original distribution was 80% good observations, 20% bad observations.

4.3.2 Re-sampling setup and performance metrics

In order for the percentage reduction in the bad observations, in each data set, to be relatively compared, the Bene1 set, Australian credit and the Behavioural Scoring set have first been altered to give a 70/30 class distribution. This was done by either under-sampling the bad observations (from a total of 1041 bad observations in the Bene1 data set, only 892 observations have been used; and from a total of 307 bad observations in the Australian credit data set, only 164 observations have been used) or under-sampling the good observations in the behavioural scoring data set, (from a total of 1436 good observations, only 838 observations have been used).

For this empirical study, the class of defaulters in each of the data sets was artificially reduced, by a factor of 5% up to 95%, so as to create a larger difference in class distribution. As a result of this reduction, six data sets were created from each of the five original data sets. For this empirical study our focus is on the performance of classification techniques on data sets with a large class imbalance. Therefore detailed results will only be presented for the data set with the original 70/30 split, as a benchmark, and data sets with 85%, 90% and 95% splits. By doing so, it is possible to identify whether techniques are adversely affected in the prediction of the target variable when there is a substantially lower number of observations in one of the classes. The performance criterion chosen to measure this effect is the area under the receiver operator characteristic curve (AUC) statistic as proposed by Baesens et al., (2003).

The receiver operating characteristic curve (ROC) is a two-dimensional graphical illustration of the trade-off between the true positive rate (sensitivity) and false positive rate (1-specificity). The ROC curve illustrates the behaviour of a classifier without having to take into consideration the class distribution or misclassification cost. In order to compare the ROC curves of different classifiers, the area under the receiver operating characteristic curve (AUC) must be computed. The AUC statistic is similar to the Gini coefficient which is equal to $2\times(AUC-0.5)$. An example of an ROC curve is depicted in FIGURE 4.1:

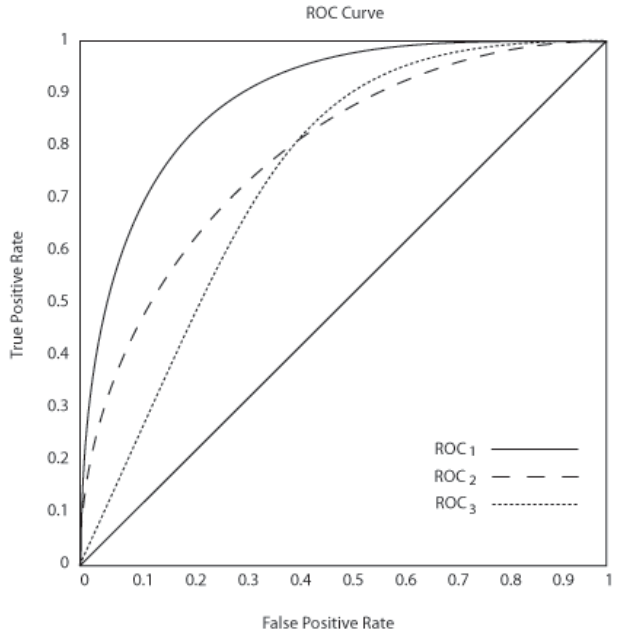


FIGURE 4.1: Example ROC Curve

The diagonal line represents the trade-off between the sensitivity and (1-specificity) for a random model, and has an AUC of 0.5. For a well performing classifier the ROC curve needs to be as far to the top left-hand corner as possible. In the example shown in FIGURE 4.1, the classifier that performs the best is that corresponding to the ROC_1 curve.

4.3.3 k-fold cross validation

For each of the techniques applied in this study a 10-fold cross validation (CV) method was applied during the modelling stage to add validity to the techniques built on the imbalanced data sets. The number of folds was selected as 10 due to the computational time for each of the different techniques over each of the data set splits. Although we would prefer a larger number of folds to reduce the bias of the true error rate estimator 10 was deemed sufficiently large for this empirical study. For the following techniques:

- Linear Discriminant Analysis (LDA)
- Quadratic Discriminant Analysis (QDA)

- Logistic Regression (LOG)
- Neural Networks (NN)
- K-nearest neighbours (k-NN)
- Gradient Boosting

this was achieved through the implementation of the group processing facility and the data transformation node in SAS Enterprise Miner. An example of the setup is presented in FIGURE 4.2:

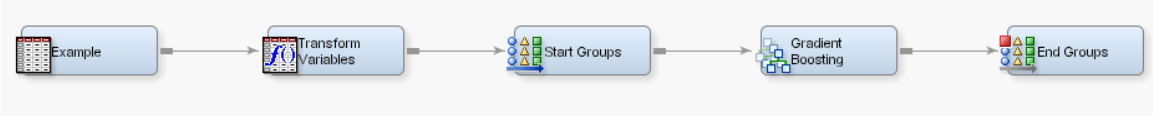


FIGURE 4.2: Example setup of k-fold cross validation

The data transformation node is required to create a random segmentation ID for the data for the k-fold groups to be used as cross validation indicators in the group processing loop. The formulation used to compute this is displayed in FIGURE 4.3:

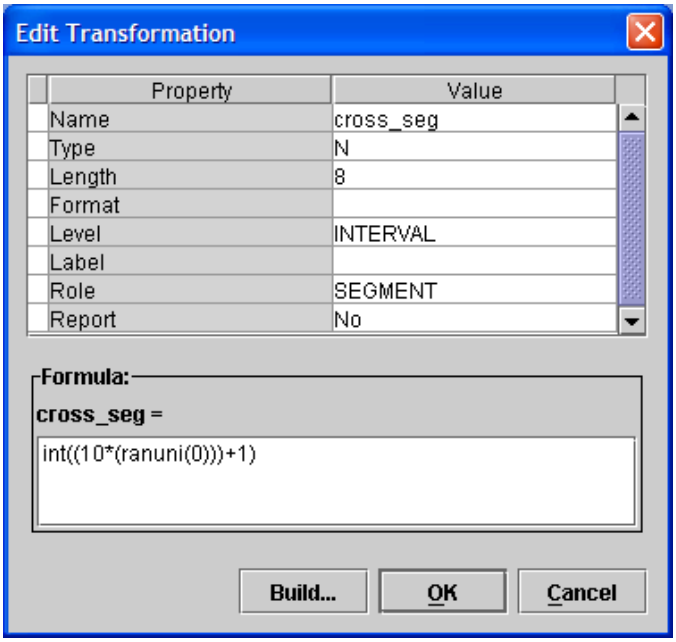


FIGURE 4.3: Transformation node in EM

(For the Random Forests and C4.5 techniques a 10-fold cross validation approach was also applied using the cross-validation option in Weka. A 10-fold cross validation

approach was also applied in the LS-SVMlab Matlab toolbox in Matlab for the LS-SVM classifier.)

Each classifier is then trained k times (k=10) using nine folds for training purposes and the reaming fold for evaluation (validation). A performance estimate for the classifier can then be determined by averaging the 10-validation estimates determined through the 10 runs of the cross validation. As mentioned in Kohavi R, (1995) common values used for k are 5 and 10 (we select 10 here in this study). Cross validation is often used to assess the performance of classification techniques on small data sets, due to the loss of potential data in the modelling process with a training/test set split. Hence why cross validation has been chosen in this instance. (Baesens, B 2003a).

4.3.4 Parameter tuning and input selection

The linear discriminant analysis (LDA), quadratic discriminant analysis (QDA) and logistic regression (LOG) classification techniques require no parameter tuning. The LOG model was built in SAS using proc logistic and using a stepwise variable selection method. Both the LDA and QDA techniques were run in SAS using proc discrim. Before all the techniques were run, dummy variables were created for the categorical variables. The AUC statistic was computed using the ROC macro by De Long et al (1988), which is available from the SAS website (http://support.sas.com/kb/25/017.html).
For the LS-SVM classifier, a linear kernel was chosen and a grid search mechanism was used to tune the hyper-parameters. For the LS-SVM, the LS-SVMlab Matlab toolbox developed by Suykens et al (2002) was used.
The NN classifiers were trained after selecting the best performing number of hidden neurons based on a validation set. The neural networks were trained in SAS Enterprise Miner using a logistic hidden and target layer activation function with the remaining EM default architecture in place (i.e. Weight Decay equal to 0, Normal randomisation distribution for random initial weights and perturbations).

The confidence level for the pruning strategy of C4.5 was varied from 0.01 to 0.5, and the most appropriate value was selected for each data set based on validation set performance. The tree was built using the Weka (Witten & Frank, 2005) package.
Two parameters have to be set for the Random Forests technique: these are the number of trees and the number of attributes used to grow each tree. A range of [10, 50, 100, 250, 500, 1000] trees has been assessed, as well as three different settings for the number of randomly selected attributes per tree $\left([0.5,1,2].\sqrt{n}\right)$, whereby n denotes the number of attributes within the respective data set (Breiman, 2001). As with the C4.5 algorithm, Random Forests were also trained in Weka (Witten & Frank, 2005), using 10-fold cross-validation for tuning the parameters.
The k-Nearest Neighbours technique was applied for both k=10 and k=100, using the Weka (Witten & Frank, 2005) IBk classifier. These values of k have been selected due to their previous use in the literature (e.g. Baesens et al 2003, Chatterjee & Barcun, 1970, West, 2000). For the Gradient Boosting classifier a partitioning algorithm was used as proposed by Friedman (2001). The number of iterations was varied in the range [10, 50, 100, 250, 500, 1000], with a maximum branch size of two selected for the splitting rule (Friedman, 2001). The gradient boosting node in SAS Enterprise Miner was used to run this technique.

4.3.5 Statistical comparison of classifiers

We used Friedman's test (Friedman, 1940) to compare the AUCs of the different classifiers. The Friedman test statistic is based on the average ranked (AR) performances of the classification techniques on each data set, and is calculated as follows:

$$\chi_F^2 = \frac{12D}{K(K+1)}\left[\sum_{j=1}^{K} AR_j^2 - \frac{K(K+1)^2}{4}\right] \text{ where } AR_j = \frac{1}{D}\sum_{i=1}^{D} r_i^j \qquad (4.1)$$

In (4.1), *D* denotes the number of data sets used in the study, *K* is the total number of classifiers and r_i^j is the rank of classifier *j* on data set *i*. χ_F^2 is distributed according to the Chi-square distribution with $K-1$ degrees of freedom. If the value of χ_F^2 is large enough, then the null hypothesis that there is no difference between the techniques can be rejected. The Friedman statistic is well suited for this type of data analysis as it is less susceptible to outliers (Friedman, 1940).

The post-hoc Nemenyi test (Nemenyi, 1963) is applied to report any significant differences between individual classifiers. The Nemenyi post-hoc test states that the performances of two or more classifiers are significantly different if their average ranks differ by at least the critical difference (CD), given by:

$$CD = q_{\alpha,\infty,K}\sqrt{\frac{K(K+1)}{12D}} \tag{4.2}$$

In this formula, the value $q_{\alpha,\infty,K}$ is based on the studentized range statistic (Nemenyi, 1963).

Finally, the results from Friedman's statistic and the Nemenyi post-hoc tests are displayed using a modified version of Demšar's (Demšar, 2006) significance diagrams (Lessmann et al., 2008). These diagrams display the ranked performances of the classification techniques along with the critical difference to clearly show any techniques which are significantly different to the best performing classifiers.

4.4 Results and discussion

The table on the following page (TABLE 4.3) reports the AUCs of all ten classifiers on the five credit scoring data sets at varying degrees of class imbalance (calculated by averaging the 10-validation estimates determined through the 10 runs of the cross validation for each classifier). For each level of imbalance, the Friedman test statistic and corresponding p-value is shown. As these were all significant ($p<0.005$), the post-hoc Nemenyi test procedure was then applied to each class distribution. The technique achieving the highest AUC on each data set is underlined as well as the overall highest ranked technique. TABLE 4.3 shows that the gradient boosting algorithm has the highest Friedman score (average rank (AR)) on two of the five different percentage class splits. However at the extreme class split (95% good, 5% bad) Random Forests provides the best average ranking across the five data sets. (N.B. example residual plots for the Gradient Boosting and Random Forest classifiers are located in Appendix A2 of this thesis).

In the majority of the class splits, the AR of the QDA and Lin LS-SVM classifiers are statistically worse than the AR of the Random Forests classifier at the 5% critical difference level ($\alpha = 0.05$), as shown in the significance diagrams included next. Note that, even though the differences between the classifiers are small, it is important to note that in a credit scoring context, an increase in the discrimination ability of even a fraction of a percent may translate into significant future savings (Henley & Hand, 1997).

	30% bad Friedman test statistic = 32.36 (p<0.005)						15% bad Friedman test statistic = 30.54 (p<0.005)						10% bad Friedman test statistic = 27.56 (p<0.005)					
	Bene1	Bene2	Germ	Aus	Behav	AR	Bene1	Bene2	Germ	Aus	Behav	AR	Bene1	Bene2	Germ	Aus	Behav	AR
LOG	80.1	77.2	76.5	91.0	65.0	5.6	79.2	77.8	75.3	91.9	67.9	5.1	78.3	79.2	76.7	85.3	65.6	4
C4.5	72.5	72.1	71.5	91.5	61.8	9	69.8	61.2	66.3	92.8	62.3	7.6	65.2	65.2	66.3	92.1	51.2	8.1
NN	79.4	77.9	73.2	91.9	72.3	5.4	76.2	78.0	69.5	91.9	70.2	5.9	76.3	77.6	72.4	90.1	69.1	5.3
Gradient Boosting	78.0	82.1	77.1	94.2	72.3	3.7	79.6	81.2	75.9	95.1	70.3	2.2	77.9	79.3	75.2	94.1	64.0	3.4
LDA	77.2	77.9	80.0	95.1	74.9	3.6	78.9	77.9	77.0	93.7	76.6	3.2	78.2	78.0	75.0	94.2	69.3	2.9
QDA	74.2	74.2	72.1	89.6	64.0	8.4	68.5	73.2	61.2	71.0	58.5	9.2	66.3	72.0	54.3	85.2	53.2	8.6
Random Forests	78.7	78.2	79.1	93.5	76.3	3.2	77.5	79.3	77.2	93.9	77.2	2.4	78.4	77.6	78.2	94.2	75.2	2
k-NN10	77.2	72.0	76.3	92.6	61.6	7.5	76.3	67.2	72.1	90.6	60.2	7.6	70.6	65.3	69.1	91.3	57.2	7.2
k-NN100	74.6	73.0	78.2	92.0	57.1	7.2	74.0	73.5	78.3	92.7	62.8	5.2	74.9	73.2	78.3	92.1	62.1	4.7
Lin LS-SVM	79.8	81.0	81.2	96.1	81.9	1.4	52.0	57.8	74.6	92.0	85.2	6.6	52.0	53.2	74.3	90.0	0.5	8.8

	5% bad Friedman test statistic = 28.32 (p<0.005)					
	Bene1	Bene2	Germ	Aus	Behav	AR
LOG	74.8	76.2	76.1	62.3	53.2	5.1
C4.5	59.3	65.0	57.8	75.4	55.3	7.6
NN	69.3	70.8	68.3	90.2	64.3	4.9
Gradient Boosting	72.1	79.0	76.2	93.2	53.4	3.2
LDA	73.8	77.2	74.2	92.6	63.2	3.2
QDA	65.3	70.2	52.3	60.1	50.6	9
Random Forests	73.5	76.3	76.3	93.3	63.0	2.4
k-NN10	66.2	63.2	68.3	88.9	53.2	7.2
k-NN100	76.0	74.2	74.9	93.1	60.2	3.4
Lin LS-SVM	52.3	53.2	53.6	86.3	51.0	9

TABLE 4.3: AUC results on test set data sets

The following significance diagrams display the AUC performance ranks of the classifiers, along with Nemenyi's critical difference (CD) tail. The CD value for all the following diagrams is equal to 6.06. Each diagram shows the classification techniques listed in ascending order of ranked performance on the y-axis, and the classifier's mean rank across all five data sets displayed on the x-axis. Two vertical dashed lines have been inserted to clearly identify the end of the best performing classifier's tail and the start of the next significantly different classifier.

The first significance diagram (see FIGURE 4.4) displays the average rank of the classifiers at the original class distribution of a 70% good, 30% bad split:

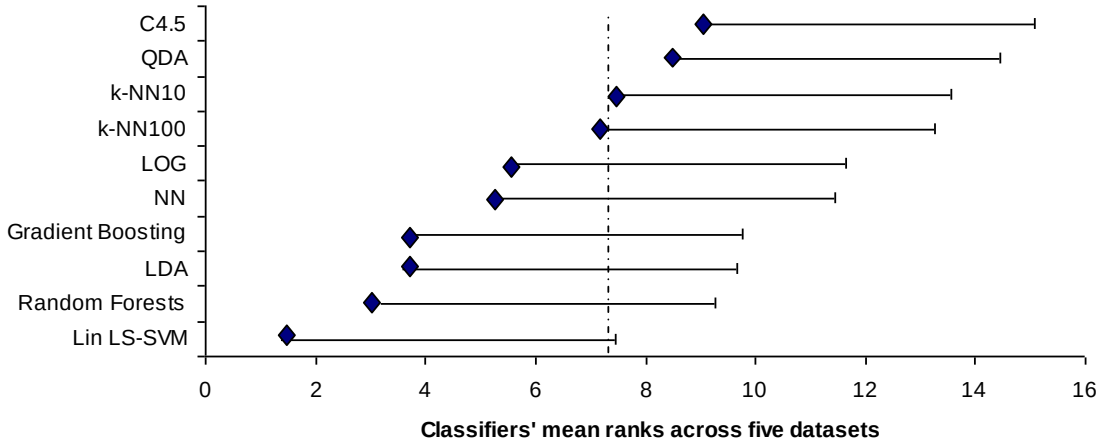


FIGURE 4.4: AR comparison at a 70/30 percentage split of good/bad observations

At this original 70/30 percentage split, the Linear LS-SVM is the best performing classification technique with an AR value of 1.4. This diagram clearly shows that the k-NN10, QDA and C4.5 techniques perform significantly worse than the best performing classifier with values of 7.5, 8.4 and 9.0 respectively.

The following significance diagram displays the average rank of the classifiers at an 85% good, 15% bad class split:

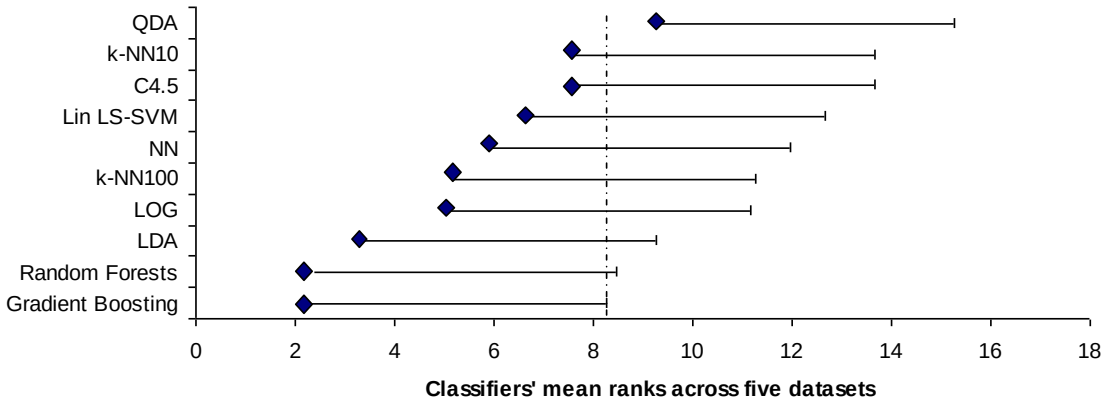


FIGURE 4.5: AR comparison at an 85/15 percentage split of good/bad observations

At the level where only 15% of the data sets are bad observations, it is shown in the significance diagram that Gradient Boosting becomes the best performing classifier (see FIGURE 4.5). The Gradient Boosting classifier performs significantly better than the quadratic discriminant analysis (QDA) classifier. From these findings we can make a preliminary assumption that when a larger class imbalance is present, the QDA classifier remains significantly different to the Gradient Boosting classifier. All the other techniques used are not significantly different.

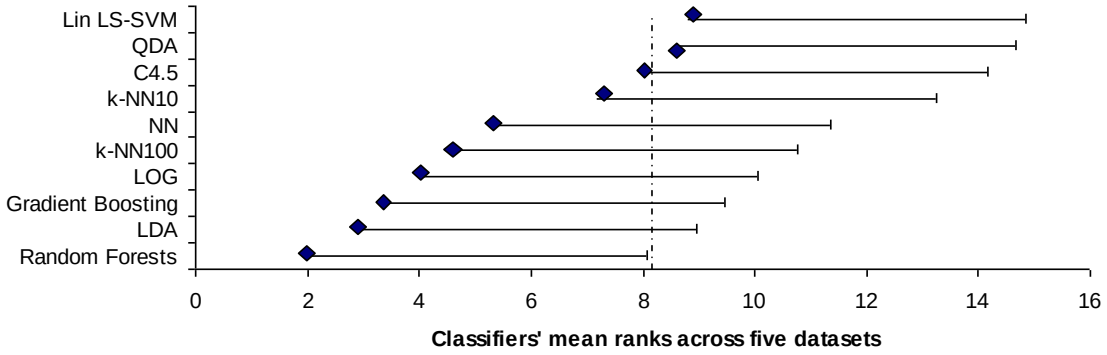


FIGURE 4.6: AR comparison at a 90/10 percentage split of good/bad observations

At a 90% good, 10% bad class split the significance diagram shown in FIGURE 4.6 indicates that the Lin LS-SVM and QDA algorithms are significantly worse than the Random Forests classifier. It can be noted that the Linear LS-SVM classifier is progressively becoming less powerful as a large class imbalance is present.

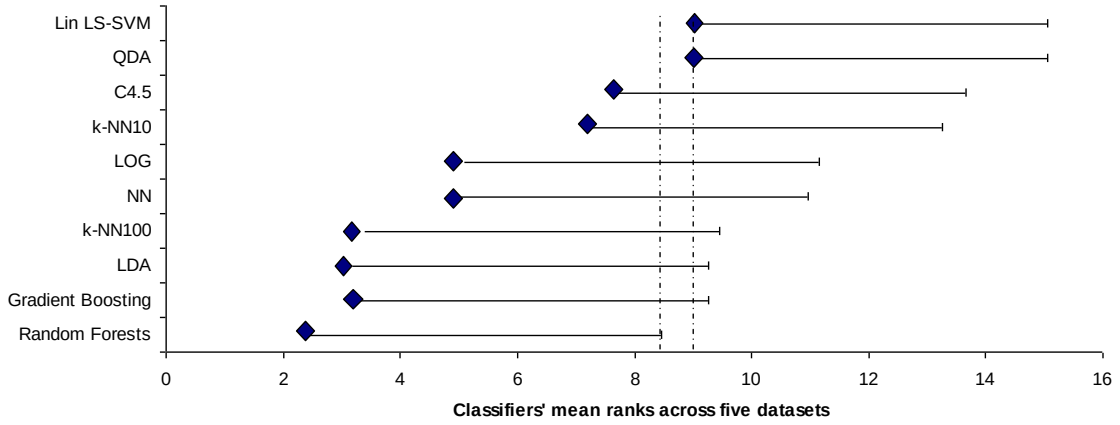


FIGURE 4.7: AR comparison at a 95/5 percentage split of good/bad observations

At a 95% good, 5% bad class split the significance diagram shown in FIGURE 4.7 indicates that the Linear LS-SVM and QDA classifiers now becomes significantly worse than the random forests classifier. This indicates that, as with the previous class split (FIGURE 4.6), the Linear LS-SVM classifier progressively becomes less powerful as a large class imbalance is present.

In summary, when considering the AUC performance measures, it can be concluded that the gradient boosting and random forest classifiers yield a very good performance at extreme levels of class imbalance, whereas the Lin LS-SVM sees a reduction in performance as a larger class imbalance is introduced. However, the simpler, linear classification techniques such as LDA and LOG also give a relatively good performance, which is not significantly different from that of the gradient boosting and random forest classifiers. This finding seems to confirm the suggestion made in (Baesens et al., 2003) that most credit scoring data sets are only weakly non-linear. The findings presented in this study show that, whereas in Yao, (2009) SVM's were shown to be the best performing classifier, at large class imbalances SVM's lose their predictive capabilities. Therefore the findings presented in this chapter agree with past analysis (Baesens et al., 2003, Yao, 2009 and Kennedy et al., 2011), but with the caveat that as a larger class imbalance is present some techniques, in particular SVM's do not perform as well. It is

also shown here that techniques such as QDA, C4.5 and k-NN10 perform significantly worse than the best performing classifiers at varying percentage reductions. The majority of classification techniques yielded classification performances that are quite competitive with each other.

4.5 Conclusions and recommendations for further work

In this comparative study we have looked at a number of credit scoring techniques, and studied their performance over various class distributions in five real-life credit data sets. Two techniques that have yet to be fully researched in the context of credit scoring, i.e. Gradient Boosting and Random Forests, were also chosen to give a broader review of the techniques available. The classification power of these techniques was assessed based on the area under the receiver operating characteristic curve (AUC). Friedman's test and Nemenyi's post-hoc tests were then applied to determine whether the differences between the average ranked performances of the AUCs were statistically significant. Finally, these significance results were visualised using significance diagrams for each of the various class distributions analysed.

The results of these experiments show that the Gradient Boosting and Random Forest classifiers performed well in dealing with samples where a large class imbalance was present. It does appear that in extreme cases the ability of random forests and gradient boosting to concentrate on 'local' features in the imbalanced data is useful. The most commonly used credit scoring techniques, linear discriminant analysis (LDA) and logistic regression (LOG), gave results that were reasonably competitive with the more complex techniques and this competitive performance continued even when the samples became much more imbalanced. This would suggest that the currently most popular approaches are fairly robust to imbalanced class sizes. On the other hand, techniques such as QDA and C4.5 were significantly worse than the best performing classifiers. It can also be concluded that the use of a linear kernel LS-SVM would not be beneficial in the scoring of data sets where a very large class imbalance exists.

Further work that could be conducted, as a result of these findings, would be to firstly consider a stacking approach to classification through the combination of multiple techniques. Such an approach would allow a meta-learner to pick the best model to classify an observation. Secondly, another interesting extension to the research would be to apply these techniques on much larger data sets which display a wider variety of class

distributions. It would also be of interest to look into the effect of not only the percentage class distribution but also the effect of the actual number of observations in a data set. Finally, as stated in the literature review chapter (cf. Chapter 2) of this thesis, there have been several approaches already researched in the area of oversampling techniques to deal with large class imbalances. Further research into this and their effect on credit scoring model performance would be beneficial.

Chapter 5

5 Estimation of Loss Given Default (LGD)

As stated in Chapter 1, the recent introduction of the Basel II framework has had a huge impact on financial institutions, allowing them to build credit risk models for three key risk parameters: PD (Probability of Default), LGD (Loss Given Default) and EAD (Exposure at Default). To date current credit risk research has largely focused on the estimation and validation of the PD parameter. However, changes in LGD directly affect the capital of a financial institution in a linear way, unlike PD, which therefore has less of an effect on minimal capital requirements. The use of models that estimate LGD as accurately as possible are thus of crucial importance as these can translate into significant future savings.

In this chapter the estimation of LGD is analysed through the implementation of various state-of-the-art regression techniques to model and predict LGD. These include one-stage models, such as those built by ordinary least squares, beta regression, artificial neural networks, support vector machines and regression trees, as well as two-stage models which attempt to combine the benefits of multiple techniques. In total 17 regression techniques are evaluated and compared using 6 real-life retail lending data sets from major international banking institutions. These particular techniques have been selected due to either their proven ability to model LGD (e.g. OLS) or their similar applications in other fields which can be transferred to a credit risk modelling context (c.f. Literature Review). The purpose of this study is to compare the widely used OLS model (with

transformations) against novel machine learning techniques to identify whether any improvement can be made in the estimation of LGD.

It is found that much of the variance of LGD remains unexplained as the average predictive performance of the models in terms of R^2 range from 4% to 43%. Nonetheless, a trend can be observed that, non-linear techniques and in particular artificial neural networks and support vector machines yield consistently higher predictive performances over all data sets than more traditional linear techniques. Also, two-stage models built by a combination of linear and non-linear techniques are shown to have similarly good predictive power, while they offer the added advantage of having a comprehensible linear model component.[1]

The remainder of this chapter is organised as follows. Section 5.2 gives a list overview of the examined regression techniques (a more detailed explanation of each of the techniques used in this chapter can be found in Chapter 3). Section 5.3 details several performance metrics for the evaluation and comparison of the regression models listed in the previous section. Section 5.4 details the data sets used and the experimental set-up implemented in this study. The penultimate section 5.5 displays the experimental results from this study, and finally section 5.6 concludes this chapter.

[1] **Nota bene: A larger version of this study was conducted as a collaborative study with the University College Ghent. Only the work contributed by the author of this thesis is presented in this chapter, except for the LS-SVM calculations which were conducted by my colleague Gert Loterman.**

5.1 Introduction

A detailed background and introduction to the topic of Loss Given Default (LGD) along with motivations for the work can be found in Chapter 1 of this thesis.

5.2 Overview of regression techniques

This study comprises both one-stage and two-stage techniques. One stage techniques can be divided into linear and non-linear techniques. Linear techniques model the dependent variable as a linear function of the independent variables while non-linear techniques fit a non-linear model to a data set. Two stage models are a combination of the aforementioned one-stage models.

The regression techniques used in this chapter comprise of both linear and non-linear techniques, and combinations of the two. A full description of these techniques can be found in Chapter 3.

Regression Techniques
Linear

Ordinary Least Squares (OLS)

Ordinary least squares regression (Draper & Smith, 1998) is the most common technique to find optimal parameters to fit a linear model to a data set. OLS estimation produces a linear regression model that minimises the sum of squared residuals for the data set.

Ordinary Least Squares with Beta transformation (B-OLS)

Before estimating an OLS model, Beta transformation/OLS (Gupton & Stein, 2002) fits a Beta distribution to the dependent variable (LGD) based on which that variable is transformed to better meet the OLS normality assumption.

Ordinary Least Squares with Box-Cox transformation (BC-OLS)

Box-Cox transformation/OLS (Box & Cox, 1964) selects an instance of a family of power transformations to improve the normality of the dependent variable.

Beta Regression (BR)

Beta Regression (Smithson & Verkuilen, 2006) uses maximum likelihood estimation to

produce a generalised linear model variant that allows for a dependent variable that is beta-distributed conditional on the input variables.

Non-linear

Regression trees (RT)

Regression Tree, sometimes referred to as classification and regression trees (CART), (Breiman, et al. 1984) algorithms produce a decision tree for the dependent variable by recursively partitioning the input space based on a splitting criterion, e.g. weighted reduction in within-node variance.

Least Squares Support Vector Machines (LSSVM)

Least Squares Support Vector Machine (Suykens, et al. 2002, Vapnik, 1995, Wang & Hu, 2005) regression implicitly maps the input space to a kernel-induced high-dimensional feature space in which a linear relationship is fitted.

Artificial Neural Networks (ANN)

Artificial Neural Networks (Bi & Bennet, 2003) produce an output value by feeding inputs through a network whose subsequent nodes apply some chosen activation function to a weighted sum of incoming values. The type of ANN considered in this study is the popular multilayer perceptron (MLP).

Log + (non-)linear

LOG+OLS, B-OLS, BC-OLS & BR

This class of two-stage (mixture) modelling approaches (Matuszyk, et al. 2010) uses logistic regression (see e.g. Hosmer & Stanley, 2000) to first estimate the probability of LGD ending up in the peak at 0 (i.e. $LGD \leq 0$) or to the right of it (i.e. $LGD > 0$). A second-stage (non-)linear regression model is built using only the observations for which $LGD > 0$. An LGD estimate is then produced by weighting the average LGD in the peak and the estimate produced by the second-stage model by their respective probabilities.

Linear + non-linear	*OLS+RT, LSSVM & ANN* The purpose of this two-stage technique (Van Gestel, et al. 2005) is to combine the good comprehensibility of OLS with the predictive power of a non-linear regression technique. In a first stage, a linear model is built using OLS. In a second stage, the residuals of this linear model estimated with a non-linear regression model. This estimate for the residual is then added to the OLS estimate to obtain a more accurate prediction for LGD. Although the concept of a two stage approach has been used before (Van Gestel, et al. 2005) it was only applied for an SVM model. This study therefore contributes the findings of an RT and ANN two-stage application as well.

TABLE 5.1: List of regression techniques

5.3 Performance metrics

Performance metrics evaluate to which degree the predictions $f(\mathbf{x}_i)$ differ from the observations y_i of the dependent variable, LGD. Each of the following metrics, listed in TABLE 5.2, has its own method to express the predictive performance of a model as a quantitative value. The second and third columns of the table show the metric values for respectively the worst and best possible prediction performance[2]. The final column shows whether the metric measures calibration or discrimination (Van Gestel & Baesens, 2009). Calibration indicates how close the predictive values are with the observed values whereas discrimination refers to the ability to provide an ordinal ranking of the dependent variable considered. A good ranking does not necessarily imply a good calibration.

Metric	Worst	Best	Measure
RMSE	$+\infty$	0	Calibration
MAE	$+\infty$	0	Calibration
AUC	0.5	1	Discrimination
AOC	$+\infty$	0	Calibration
R^2	0	1	Calibration
r	0	1	Discrimination
ρ	0	1	Discrimination
τ	0	1	Discrimination

TABLE 5.2: Performance Metrics

[2] Note that the R^2 measure defined here could possibly lie outside the [0,1] interval when applied to non-OLS models. Although alternative generalised goodness-of-fit measures have been put forward for evaluating various non-linear models (see e.g. Nagelkerke, 1991), the measure defined in TABLE 5.2 has the advantage that it is widely used and can be calculated for all techniques.

5.3.1 Root Mean Squared Error (RMSE)

RMSE (see for example, Draper & Smith, 1998) is defined as the square root of the average of the squared difference between predictions and observations:

$$RMSE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(f\left(\mathbf{x}_i\right)-y_i\right)^2} \tag{5.1}$$

RMSE has the same units as the independent variable being predicted. Since residuals are squared, this metric heavily weights outliers. The smaller the value of RMSE the better the prediction, with 0 being a perfect prediction.

5.3.2 Mean Absolute Error (MAE)

MAE (see for example, Draper & Smith, 1998) is given by the averaged absolute differences of predicted and observed values:

$$MAE = \frac{1}{l}\sum_{i=1}^{l}\left|f\left(\mathbf{x}_i\right)-y_i\right| \tag{5.2}$$

Just like RMSE, MAE has the same unit scale as the dependent variable being predicted. Unlike RMSE, MAE is not that sensitive to outliers. The metric is bound between the maximum absolute error and 0 (perfect prediction).

5.3.3 Area under the Receiver Operating Curve (AUC)

ROC curves are normally used for the assessment of binary classification techniques (see for example, Fawcett, 2006). It is however used in this context to measure how good the regression technique is in distinguishing high values from low values of the dependent variable. To build the ROC curve, the observed values are first classified into high and low classes using the mean $\bar{y}$ of the training set as reference.

5.3.4 Area over the Regression Error Characteristic curves (AOC)

REC curves (Bi & Bennet, 2003) generalise ROC curves for regression. The AOC curve plots the error tolerance on the x-axis versus the percentage of points predicted within the tolerance (or accuracy) on the y-axis (FIGURE 5.1). The resulting curve estimates the cumulative distribution function of the squared error. The area over the REC curve (AOC) is an estimate of the predictive power of the technique. The metric is bound between 0 (perfect prediction) and the maximum squared error.

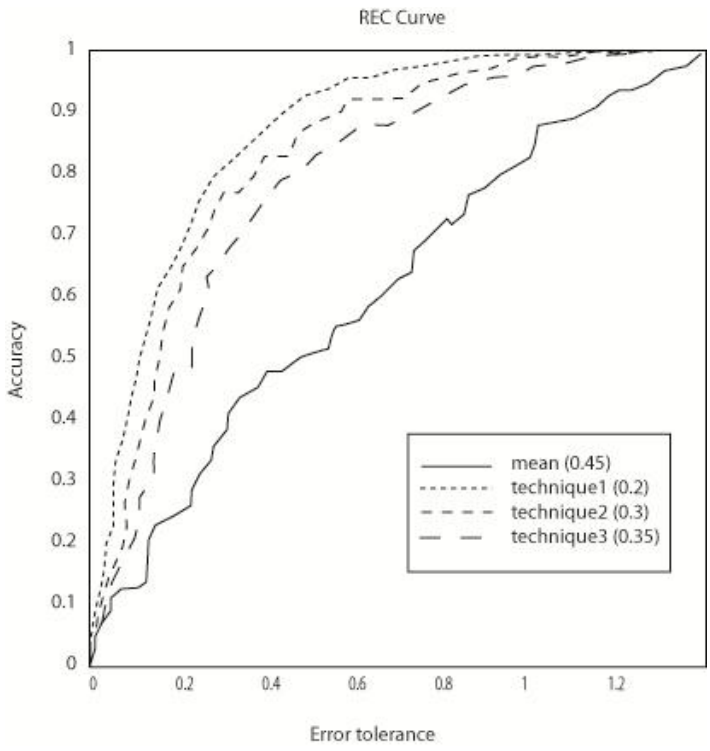


FIGURE 5.1: Example REC Curve

5.3.5 Coefficient of Determination (R^2)

The Coefficient of Determination R^2 (see for example, Draper & Smith, 1998) can be defined as 1 minus the fraction of the residual sum of squares to the total sum of squares:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}} \tag{5.3}$$

Where $SS_{err} = \sum_{i=1}^{l} \left(y_i - f\left(\mathbf{x}_i\right) \right)^2$, $SS_{tot} = \sum_{i=1}^{l} \left(y_i - \bar{y} \right)^2$ and $\bar{y}$ is the mean of the observed values. Since the second term in the formula can be seen as the fraction of unexplained variance, the R^2 can be interpreted as the fraction of explained variance. Although R^2 is usually expressed as a number on a scale from 0 to 1, R^2 can yield negative values when the model predictions are worse than using the mean $\bar{y}$ from the training set as prediction. Although alternative generalised goodness-of-fit measures have been put forward for evaluating various non-linear models (see e.g. Nagelkerke, 1991), R^2 has the advantage that it is widely used and can be calculated for all techniques.

5.3.6 Pearson's Correlation Coefficient (r)

Pearson's *r* (see e.g. Cohen, et al. 2002) is defined as the sum of the products of the standard scores of the observed and predicted values divided by the degrees of freedom:

$$r = \frac{1}{l-1} \sum_{i=1}^{l} \left(\frac{y_i - \bar{y}}{s_y} \right) \left(\frac{f\left(\mathbf{x}_i\right) - \bar{f}}{s_f} \right) \tag{5.4}$$

with $\bar{y}$ and $\bar{f}$ the mean and s_y and s_f the standard deviation of respectively the observations and predictions. Pearson's *r* can take values between -1 (perfect negative correlation) and +1 (perfect positive correlation) with 0 meaning no correlation at all.

5.3.7 Spearman's Correlation Coefficient (ρ)

Spearman's ρ (see e.g. Cohen, et al. 2002) is defined as Pearson's *r* applied to the rankings of predicted and observed values. If there are no (or very few) tied ranks however, it is common to use the equivalent formula:

$$\rho = 1 - \frac{6\sum_{i=1}^{l} d_i^2}{l\left(l^2 - 1\right)} \tag{5.5}$$

where d_i is the difference between the ranks of observed and predicted values. Spearman's ρ can take values between -1 (perfect negative correlation) and +1 (perfect positive correlation) with 0 meaning no correlation at all.

5.3.8 Kendall's Correlation Coefficient (τ)

Kendall's τ (see e.g. Cohen, et al. 2002) measures the degree of correspondence between observed and predicted values. In other words, it measures the association of cross tabulations:

$$\tau = \frac{n_c - n_d}{\frac{1}{2} l\left(l - 1\right)} \tag{5.6}$$

where n_c is the number of concordant pairs and n_d is the number of discordant pairs. A pair of observations $\{i, k\}$ is said to be concordant when there is no tie in either observed or predicted LGD (i.e. $y_i \neq y_k$, $f\left(\mathbf{x}_i\right) \neq f\left(\mathbf{x}_k\right)$), and if $\operatorname{sgn}\left(f\left(\mathbf{x}_k\right) - f\left(\mathbf{x}_i\right)\right) = \operatorname{sgn}\left(y_k - y_i\right)$, where $i, k = 1, \ldots, l\left(i \neq k\right)$. Similarly, it is said to be discordant if there is no tie and if $\operatorname{sgn}\left(f\left(\mathbf{x}_k\right) - f\left(\mathbf{x}_i\right)\right) = -\operatorname{sgn}\left(y_k - y_i\right)$. Kendall's τ can take values between -1 (perfect negative correlation) and +1 (perfect positive correlation) with 0 meaning no correlation is present.

5.4 Experimental set-up and data sets

In this section the characteristics of the data sets are described as well as the experimental benchmarking framework to assess the predictive performance of the regression techniques. Further, a description of a technique's parameter setting and tuning is given where required.

5.4.1 Data set characteristics

TABLE 5.3 displays the characteristics of 6 real-life lending LGD data sets from a series of financial institutions, each of which contains loan-level data about defaulted loans and their resulting losses. The number of data set entries varies from a few thousands to just under 120,000 observations. The number of available input variables ranges from 12 to 44. The types of loan data set included are personal loans, corporate loans, revolving credit and mortgage loans. The empirical distribution of LGD values observed in each of the data sets is displayed in FIGURE 5.2. Note that the LGD distribution in consumer lending often contains one or two spikes around $LGD = 0$ (in which case there was a full recovery) and/or $LGD = 1$ (no recovery). Also, a number of data sets include some LGD values that are negative (e.g., because of penalties paid, gains in collateral sales, etc.) or larger than 1 (e.g., due to additional collection costs incurred); in other data sets, values outside the unit interval were truncated to 0 or 1 by the banks themselves. Importantly LGD does not display a normal distribution in any of these data sets.

Data set	Type	Inputs	Data set size	Training set size	Test set size
BANK 1	Personal loans	44	47,853	31,905	15,948
BANK 2	Mortgage loans	18	119,211	79,479	39,732
BANK 3	Mortgage loans	14	3,351	2,232	1,119
BANK 4	Revolving credit	12	7,889	5,260	2,629
BANK 5	Mortgage loans	35	4,097	2,733	1,364
BANK 6	Corporate loans	21	4,276	2,851	1,425

TABLE 5.3: Data set characteristics of real-life LGD data

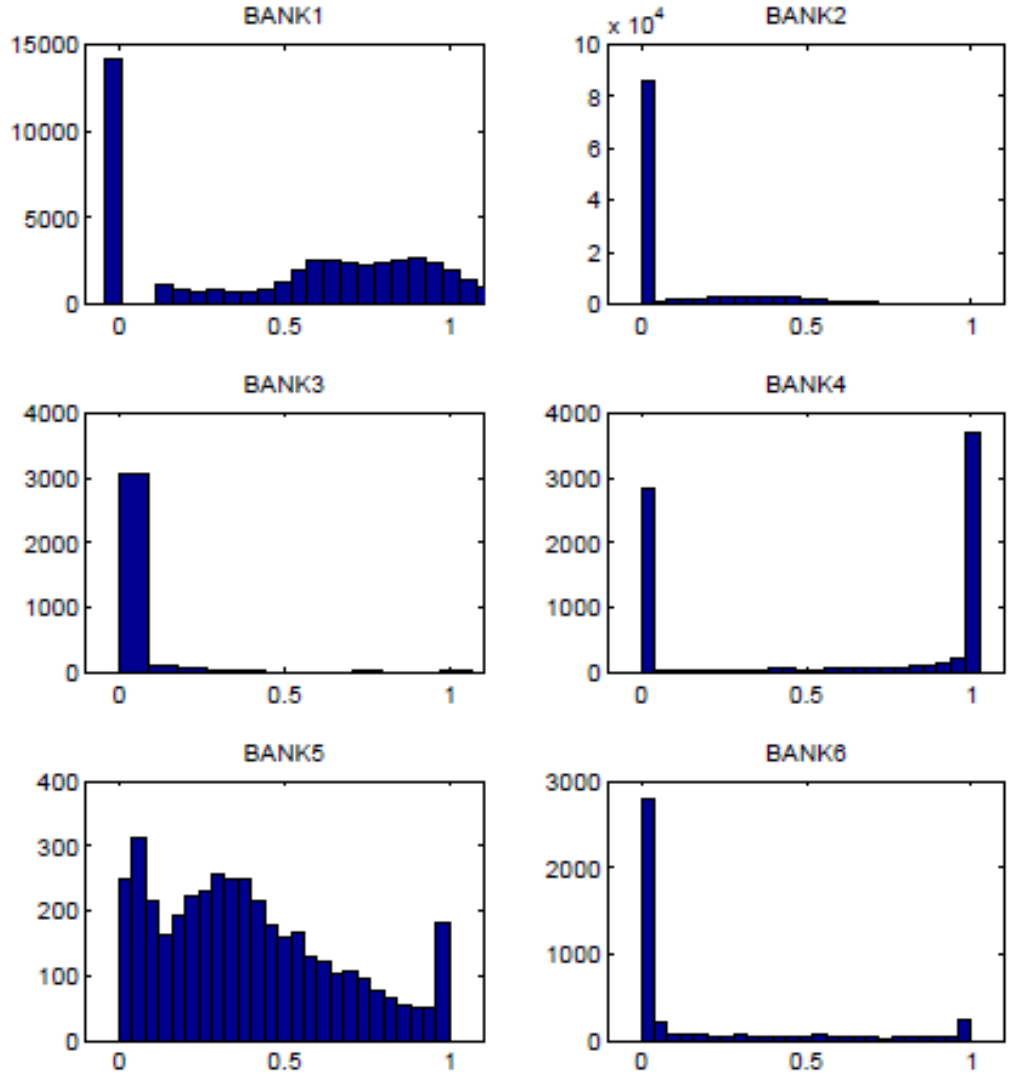


FIGURE 5.2: LGD distributions of real-life LGD data sets

5.4.2 Experimental set-up

First, each data set is randomly shuffled and divided into two thirds training set and one third test set. The training set is used to build the models while the test set is solely used to assess the predictive performance of these models. Where required, continuous independent variables are standardised with the sample mean and standard deviation of the training set, nominal independent variables are encoded with dummy variables and ordinal independent variables are encoded with thermo variables.

An input selection method is used to remove irrelevant and redundant variables from the data set, with the aim of improving the performance of regression techniques. For this, a stepwise selection method is applied for building the linear models (APPENDIX A3). For computational efficiency reasons, an R^2 based filter method (Freund & Littell, 2000) is applied prior to building the non-linear models (APPENDIX A4).

After building the models, the predictive performance of each data set is measured on the test set by comparing the predictions and observations according to several performance metrics. Next, an average ranking of techniques over all data sets is generated per performance metric as well as a meta-ranking of techniques over all data sets and all performance metrics.

Finally, the regression techniques are statistically compared with each other (Demsar, 2006). A Friedman test (Friedman, 1940) is performed to test the null hypothesis that all regression techniques perform alike according to a specific performance metric, i.e., performance differences would just be due to random chance. A more detailed summary and the applied formulas can be found in the previous chapter (cf. Chapter 4.3.4).

5.4.3 Parameter settings and tuning

During model building, several techniques require parameters to be set or tuned. This section describes how these are set or tuned where appropriate.

5.4.3.1 Ordinary Least Squares (OLS)

For the OLS technique no extra parameter tuning is required.

5.4.3.2 Ordinary Least Squares with Beta transformation (B-OLS)

For the B-OLS technique no extra parameter tuning is required.

5.4.3.3 Ordinary Least Squares with Box-Cox transformation (BC-OLS)

The value of parameter *c* is set to zero. The value of the power parameter λ is varied over a chosen range (e.g. from -3 to 3 in 0.25 increments) and an optimal value is chosen based on a maximum likelihood criterion.

5.4.3.4 Beta Regression (BR)

For the BR technique no extra parameter tuning is required.

5.4.3.5 Regression Trees (RT)

For the regression tree model, the training set is further split into a training and validation set. The validation set is used to select the criterion for evaluating candidate splitting rules (i.e. variance reduction or ProbF), the depth of the tree, and the threshold p-value for the ProbF criterion. The choice of tree depth, the threshold p-value for the ProbF

criterion and criterion method was selected based on the mean squared error on the validation set.

5.4.3.6 Least Squares Support Vector Machines (LSSVM)

Although several kernels can be used, the radial basis function (RBF) kernel:

$$K\left(\mathbf{x},\mathbf{x}_i\right)=e^{-\frac{\left\|\mathbf{x}-\mathbf{x}_i\right\|^2}{2\sigma^2}} \tag{5.7}$$

with kernel parameter σ is used here because of its good overall performance for LSSVM classifiers (Baesens, et al. 2000). The hyper parameters γ and σ for LSSVM regression are tuned with 10-fold cross validation on the training data set. A grid search evaluates all possible combinations of parameters within the search space in order to find a possible optimal combination that minimises the mean squared error. The limits of the grid for the kernel parameter σ are set to $\left[0.5\cdot\sqrt{l},500\cdot\sqrt{l}\right]$ and the limits of the grid for the regularisation parameter γ are set to $\left[\frac{0.01}{n},\frac{1000}{n}\right]$ (Van Gestel, et al. 2003).

Estimating the LSSVM hyper parameters this way can be a computational burden. To tune the hyper parameters, a sample from the complete training data set is chosen as follows. First, 100 random subsets of 4000 observations are chosen. Next, the LGD distribution histogram of each subset is compared with the LGD distribution histogram of the complete training set, and the subset that best approximates the original set based on the mean squared error, is chosen.

5.4.3.7 Artificial Neural Networks (ANN)

For the ANN model, the training set is further split into a training and validation set. The validation is used to evaluate the target layer activation functions (logistic, linear, exponential, reciprocal, square, sine, cosine, tanh and arcTan) and number of hidden neurons (1-20) used in the model. The weights of the network are first randomly

initialised and then iteratively adjusted so as to minimise the mean squared error. The choice of activation function and number of hidden neurons is selected based on the mean squared error on the validation set. The hidden layer activation function is set as logistic.

5.5 Results and discussion

TABLES 5.4 to 5.9 contain the performance results obtained for all techniques on the 6 respective data sets. The best performing model according to each metric is underlined. FIGURE 5.3 displays a series of box plots for the observed distributions of performance values for the metrics *AUC*, R^2, r, ρ and τ. Similar trends can be observed across all metrics. Note that differences in type of data set, number of observations and available independent variables are the likely causes of the observed variability of actual performance levels between the 6 different data sets.

Although all performance metrics listed above are useful measures in their own right, it is common to use the coefficient of determination R^2 to compare model performance, since R^2 measures calibration and can be compared meaningfully across different data sets. As shown in FIGURE 5.3, the average R^2 of the models varies from about 4 % to 43 %. In other words, the variance in LGD that can be explained by the independent variables is consistently below 50 %, implying that most of the variance cannot be explained even with the best models. Note that although R^2 usually is a number on a scale from 0 to 1, R^2 can yield negative values for non-OLS models when the model predictions are worse than always using the mean from the training set as prediction.

Technique	MAE	RMSE	AUC	AOC	R^2	r	ρ	τ
OLS	0.3257	0.3716	0.6570	0.138	0.0972	0.3112	0.3084	0.2145
B-OLS	0.3474	0.4294	0.6580	0.1843	- 0.2060	0.2954	0.2991	0.2071
BC-OLS	0.3835	0.4579	0.5180	0.2096	- 0.3747	0.2403	0.2312	0.1602
BR	0.3356	0.3693	0.5690	0.1363	0.0546	0.2601	0.2641	0.1844
RT	0.3228	0.3732	0.5990	0.1392	0.0892	0.2997	0.2913	0.2095
LSSVM	0.3184	0.3669	0.6723	0.1346	0.1194	0.3466	0.3442	0.2444
ANN	0.3118	0.3648	0.6840	0.1331	0.1295	0.3603	0.3559	0.2524
LOG+OLS	0.3202	0.3700	0.6210	0.1366	0.1063	0.3262	0.3143	0.2214
LOG+B-OLS	0.3163	0.3750	0.6020	0.1406	0.1002	0.3166	0.3103	0.2185
LOG+BC-OLS	0.4308	0.5090	0.5040	0.2590	- 0.6946	0.2125	0.2440	0.1731
LOG+BR	0.3560	0.4142	0.5270	0.1715	0.0782	0.2797	0.2591	0.1794
LOG+RT	0.3219	0.3693	0.6160	0.1363	0.1081	0.3301	0.3212	0.2263
LOG+LSSVM	0.3191	0.3679	0.6664	0.1353	0.1150	0.3401	0.3336	0.2371
LOG+ANN	0.3174	0.3664	0.6320	0.1342	0.1221	0.3502	0.3406	0.2395
OLS+RT	0.3170	0.3681	0.6730	0.1354	0.1137	0.3382	0.3342	0.2348
OLS+LSSVM	0.3115	0.3631	0.6929	0.1317	0.1379	0.3714	0.3666	0.2596
OLS+ANN	0.3079	0.3633	0.6960	0.1318	0.1367	0.3716	0.3638	0.2581

TABLE 5.4: BANK 1 performance results

Technique	MAE	RMSE	AUC	AOC	R^2	r	ρ	τ
OLS	0.1187	0.1613	0.8100	0.0259	0.2353	0.4851	0.4890	0.3823
B-OLS	0.1058	0.1621	0.8000	0.0262	0.2273	0.4768	0.4967	0.3881
BC-OLS	0.1056	0.1623	0.7450	0.0262	0.2226	0.4718	0.4990	0.3900
BR	0.1020	0.1661	0.7300	0.0275	0.2120	0.4635	0.4857	0.3861
RT	0.0978	0.1499	0.7710	0.0224	0.3390	0.5823	0.5452	0.4357
LSSVM	0.1047	0.1518	0.8365	0.0230	0.3229	0.5690	0.5301	0.4160
ANN	0.0956	0.1472	0.8530	0.0216	0.3632	0.6029	0.5549	0.4366
LOG+OLS	0.1060	0.1622	0.7590	0.0255	0.2268	0.4838	0.5206	0.4084
LOG+B-OLS	0.1040	0.1567	0.8320	0.0245	0.2779	0.5286	0.5202	0.4083
LOG+BC-OLS	0.1034	0.1655	0.7320	0.0273	0.2124	0.4628	0.4870	0.3820
LOG+BR	0.1015	0.1688	0.7250	0.0285	0.2024	0.4529	0.4732	0.3876
LOG+RT	0.1041	0.1538	0.8360	0.0236	0.3049	0.5545	0.5254	0.4126
LOG+LSSVM	0.1031	0.1530	0.8334	0.0234	0.3121	0.5587	0.5243	0.4128
LOG+ANN	0.1011	0.1531	0.8430	0.0234	0.3109	0.5585	0.5380	0.4240
OLS+RT	0.1015	0.1506	0.8410	0.0227	0.3331	0.5786	0.5344	0.4188
OLS+LSSVM	0.1029	0.1520	0.8428	0.0230	0.3208	0.5665	0.5398	0.4241
OLS+ANN	0.0999	0.1474	0.8560	0.0217	0.3612	0.6010	0.5585	0.4398

TABLE 5.5: BANK 2 performance results

Technique	MAE	RMSE	AUC	AOC	R^2	r	ρ	τ
OLS	0.0549	0.1411	0.6460	0.0178	0.0124	0.1168	0.0965	0.0718
B-OLS	0.0348	0.1449	0.6610	0.0188	-0.0419	0.0767	0.1754	0.1361
BC-OLS	0.0340	0.1456	0.6380	0.0190	-0.0529	0.1373	0.2312	0.1765
BR	0.0883	0.1315	0.6530	0.0169	-0.1128	0.1567	0.1719	0.1323
RT	0.0482	0.1311	0.6990	0.0154	0.1477	0.3869	0.2007	0.1673
LSSVM	0.0473	0.1270	0.7441	0.0140	0.1998	0.4526	0.2085	0.1520
ANN	0.0458	0.1318	0.6000	0.0152	0.1386	0.3776	0.1482	0.1105
LOG+OLS	0.0553	0.1417	0.6010	0.0179	0.0043	0.0759	0.0701	0.0510
LOG+B-OLS	0.0392	0.1429	0.6330	0.0182	-0.0127	0.1214	0.1252	0.0923
LOG+BC-OLS	0.0349	0.1448	0.6330	0.0188	-0.0395	0.1665	0.1918	0.1426
LOG+BR	0.0569	0.1417	0.5790	0.0180	0.0043	0.0742	0.1710	0.1265
LOG+RT	0.0434	0.1297	0.7210	0.0146	0.1663	0.4553	0.1571	0.1170
LOG+LSSVM	0.0460	0.1312	0.7485	0.0151	0.1471	0.4152	0.2272	0.1676
LOG+ANN	0.0452	0.1219	0.6190	0.0133	0.2634	0.5381	0.1671	0.1242
OLS+RT	0.0540	0.1372	0.7050	0.0168	0.0660	0.2578	0.1748	0.1285
OLS+LSSVM	0.0483	0.1258	0.7416	0.0137	0.2148	0.4648	0.1869	0.1354
OLS+ANN	0.0570	0.1388	0.6730	0.0171	0.0442	0.2605	0.1369	0.1005

TABLE 5.6: BANK 3 performance results

Technique	MAE	RMSE	AUC	AOC	R^2	r	ρ	τ
OLS	0.2712	0.3479	0.8520	0.1208	0.4412	0.6643	0.5835	0.4331
B-OLS	0.2214	0.3743	0.8500	0.1396	0.3530	0.6510	0.5822	0.4321
BC-OLS	0.3185	0.4292	0.6750	0.1839	0.1478	0.5726	0.5820	0.4316
BR	0.3208	0.3777	0.8480	0.1425	0.3405	0.6527	0.5908	0.4452
RT	0.2476	0.3362	0.8480	0.1128	0.4782	0.6916	0.5919	0.4762
LSSVM	0.2428	0.3315	0.8655	0.1097	0.4924	0.7017	0.6203	0.4692
ANN	0.2393	0.3299	0.8670	0.1086	0.4974	0.7053	0.6109	0.4555
LOG+OLS	0.2577	0.3465	0.8520	0.1199	0.4455	0.6678	0.5840	0.4338
LOG+B-OLS	0.2399	0.3551	0.8500	0.1259	0.4176	0.6651	0.5801	0.4301
LOG+BC-OLS	0.2502	0.3489	0.8510	0.1215	0.4379	0.6659	0.5819	0.4322
LOG+BR	0.2738	0.3560	0.8520	0.1265	0.4147	0.6680	0.5868	0.4342
LOG+RT	0.2679	0.3621	0.8570	0.1309	0.3945	0.6656	0.5899	0.4364
LOG+LSSVM	0.2534	0.3425	0.8590	0.1172	0.4581	0.6771	0.6024	0.4541
LOG+ANN	0.2558	0.3457	0.8540	0.1184	0.4480	0.6698	0.5852	0.4348
OLS+RT	0.2628	0.3425	0.8590	0.1171	0.4582	0.6776	0.6017	0.4498
OLS+LSSVM	0.2439	0.3322	0.8656	0.1102	0.4904	0.7003	0.6211	0.4698
OLS+ANN	0.2404	0.3300	0.8710	0.1087	0.4971	0.7053	0.6195	0.4635

TABLE 5.7: BANK 4 performance results

Technique	MAE	RMSE	AUC	AOC	R^2	r	ρ	τ
OLS	0.1875	0.2375	0.7480	0.0555	0.2218	0.474	0.5192	0.3651
B-OLS	0.1861	0.2368	0.7410	0.0561	0.2263	0.5073	0.5168	0.3636
BC-OLS	0.1848	0.2373	0.7390	0.0560	0.2228	0.5014	0.5155	0.3632
BR	0.1957	0.2402	0.7240	0.0575	0.2038	0.4557	0.4811	0.3359
RT	0.1851	0.2324	0.7370	0.0538	0.2546	0.5056	0.4957	0.3888
LSSVM	0.1707	0.2198	0.7847	0.0479	0.3331	0.5794	0.5801	0.4167
ANN	0.1678	0.2173	0.7830	0.0470	0.3486	0.5964	0.5765	0.4148
LOG+OLS	0.1851	0.2336	0.7500	0.0542	0.2468	0.4975	0.5246	0.3704
LOG+B-OLS	0.1852	0.2347	0.7480	0.0548	0.2397	0.5117	0.5192	0.3658
LOG+BC-OLS	0.1833	0.2349	0.7470	0.0549	0.2388	0.5099	0.5238	0.3699
LOG+BR	0.1939	0.2395	0.7250	0.0572	0.2083	0.4568	0.4820	0.3364
LOG+RT	0.1846	0.2344	0.7380	0.0547	0.2420	0.5000	0.4903	0.3445
LOG+LSSVM	0.1708	0.2197	0.7835	0.0479	0.3340	0.5797	0.5795	0.4163
LOG+ANN	0.1689	0.2188	0.7810	0.0476	0.3396	0.5845	0.5737	0.4135
OLS+RT	0.1779	0.2320	0.7660	0.0530	0.2572	0.5357	0.5554	0.3963
OLS+LSSVM	0.1695	0.2216	0.7882	0.0485	0.3223	0.5755	0.5933	0.4279
OLS+ANN	0.1747	0.2277	0.7730	0.0510	0.2844	0.5567	0.5706	0.4086

TABLE 5.8: BANK 5 performance results

Technique	MAE	RMSE	AUC	AOC	R^2	r	ρ	τ
OLS	0.2085	0.2874	0.7180	0.0822	0.1197	0.3502	0.3032	0.2071
B-OLS	0.1783	0.3055	0.7120	0.0933	0.0933	0.3054	0.3112	0.2138
BC-OLS	0.1824	0.3149	0.7100	0.0988	0.0815	0.2855	0.3139	0.2172
BR	0.2612	0.3019	0.7090	0.0909	0.1029	0.3209	0.3138	0.2151
RT	0.2061	0.2885	0.7040	0.0829	0.1129	0.3390	0.3180	0.2482
LSSVM	0.2031	0.2812	0.7360	0.0787	0.1570	0.3964	0.3207	0.2190
ANN	0.2004	0.2860	0.7210	0.0815	0.1281	0.3619	0.2893	0.2000
LOG+OLS	0.2086	0.2876	0.7180	0.0824	0.1182	0.3479	0.3012	0.2060
LOG+B-OLS	0.1899	0.2964	0.7070	0.0875	0.0635	0.3225	0.2913	0.2000
LOG+BC-OLS	0.1863	0.3055	0.7120	0.0930	0.0963	0.3103	0.3050	0.2118
LOG+BR	0.2875	0.3204	0.7070	0.1024	-0.0946	0.3346	0.2806	0.1918
LOG+RT	0.2052	0.2890	0.6880	0.0832	0.1100	0.3348	0.3179	0.2219
LOG+LSSVM	0.2024	0.2887	0.7191	0.0829	0.1116	0.3652	0.3159	0.2190
LOG+ANN	0.2038	0.2854	0.7290	0.0811	0.1319	0.3689	0.3243	0.2216
OLS+RT	0.2066	0.2866	0.7190	0.0817	0.1244	0.3623	0.3067	0.2100
OLS+LSSVM	0.2087	0.2875	0.7180	0.0822	0.1189	0.3493	0.3030	0.2070
OLS+ANN	0.2085	0.2874	0.7190	0.0822	0.1200	0.3498	0.3049	0.2086

TABLE 5.9: BANK 6 performance results

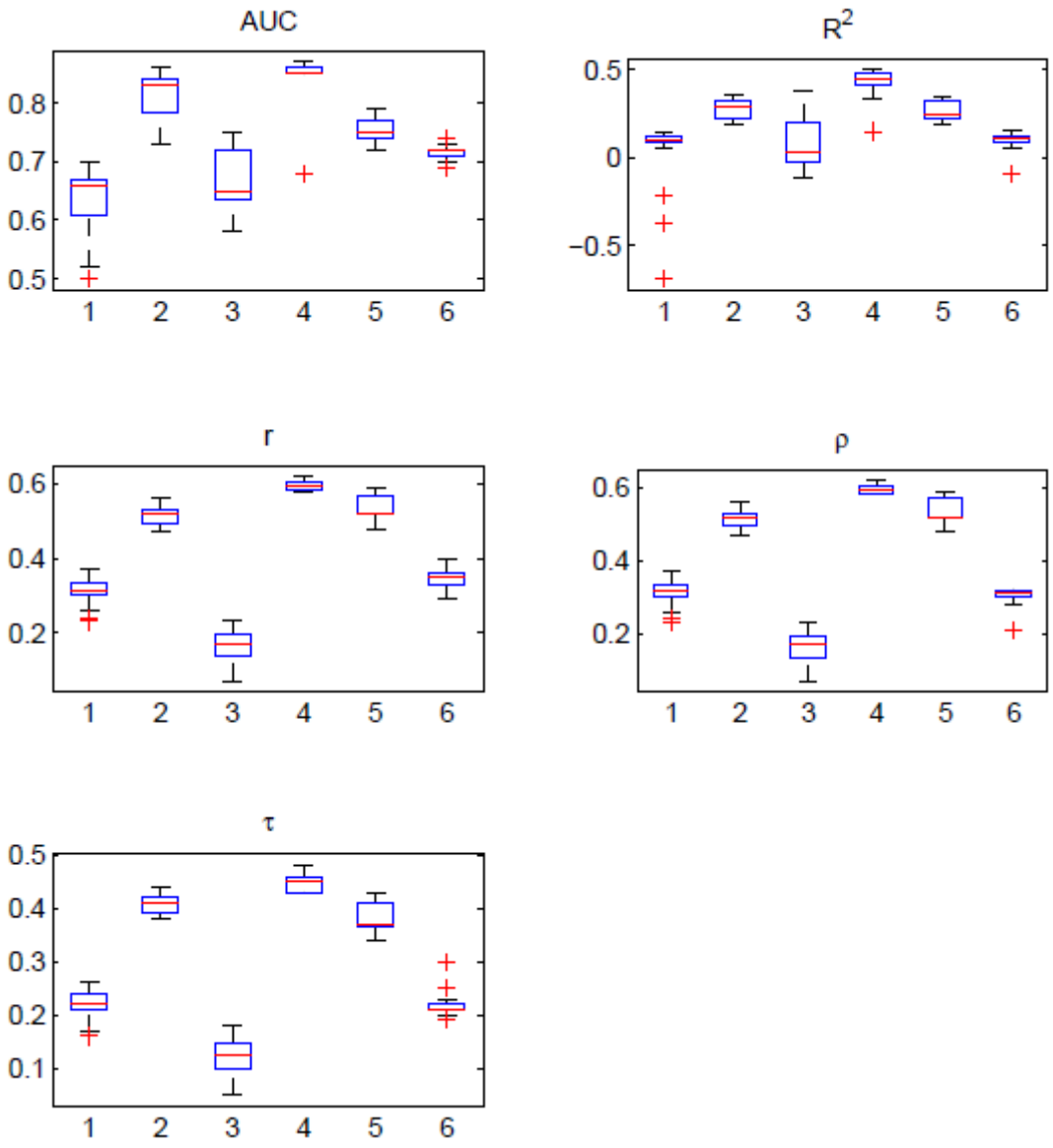


FIGURE 5.3: Comparison of predictive performances across 6 real-life retail lending data sets

The linear models that incorporate some form of transformation to the dependent variable (i.e. B-OLS, BR, BC-OLS) are shown to perform consistently worse than OLS, despite the fact that these approaches are specifically designed to cope with the violation of the OLS normality assumption. This suggest that they too have difficulties dealing with the pronounced point densities observed in LGD data sets, while they may be less efficient than OLS or they could introduce model bias if a transformation is performed prior to OLS estimation (as is the case for B-OLS and BC-OLS).

Perhaps the most striking result is that, in contrast with prior benchmarking studies on classification models for PD (Baesens, et al. 2003), non-linear models such as LSSVM and ANN significantly outperform most linear models in the prediction of LGD. This implies that the relation between LGD and the independent variables in the data sets is non-linear (as is most apparent on data set BANK3, see TABLE 5.6). Also, LSSVM and ANN generally perform better than RT. However, LSSVM and ANN result in black-box models while RT have the ability to produce comprehensible white-box models. To circumvent this disadvantage, one could try to obtain an interpretation for a well-performing black-box model by applying rule extraction techniques (Martens, et al. 2007, Martens, et al. 2009).

The performance evaluation of the class of two-stage models in which a logistic regression model is combined with a second-stage (linear or non-linear) model ($LOG+$), is less straightforward. Although a weak trend is noticeable that logistic regression combined with a linear model tends to increase the performance of the latter, it appears that logistic regression combined with a non-linear model slightly reduces the strong performance of the latter. Because the LGD distributions from BANK4, BANK5 and BANK6 also show a peak at $LGD=1$, the performance of these models could possibly be increased by slightly altering the technique. Replacing the (binary) logistic regression component by an ordinal logistic regression model distinguishing between 3 classes ($LGD \leq 0, 0 < LGD < 1, LGD \geq 1$) and then using a second-stage model for $0 < LGD < 1$ could perhaps better account for the presence of both peaks.

In contrast with the previous two-stage model, a clear trend can be observed for the combination of a linear and a non-linear model ($OLS+$). By estimating the error residual of an OLS model using a non-linear technique, the prediction performance tends to increase to somewhere very near the level of the corresponding one-stage non-linear technique. What makes these two-stage models attractive is that they have the advantage of combing the high prediction performance of non-linear regression with the comprehensibility of a linear regression component. Note that this modelling method has

also been successfully applied in a PD modelling context Van Gestel, et al. 2005, Van Gestel, et al. 2006, Van Gestel, et al. 2007).

The average ranking over all data sets according to each performance metric is listed in columns 2 to 9 of TABLE 5.10. The best performing technique for each metric is underlined and techniques that significantly perform worse than the best performing technique for that metric according to the Nemenyi's post-hoc test $(\alpha = 0.5)$ are in italic. The last column illustrates the meta-ranking (MR) as the average ranking (AR) over all data sets and over all metrics. The techniques in the table are sorted according to their meta-ranking. Additionally, columns 10 and 11 cover the meta-ranking only including respectively calibration and discrimination metrics. The best performing techniques are consistently ranked in the top according to each metric, no matter whether they measure calibration or discrimination.

The results of the Friedman test and subsequent Nemenyi's post-hoc test with significance level $\alpha = 0.05$ can be intuitively visualised using Demsar's significance diagram (Demsar, 2006). FIGURES 5.4-5.11 display the Demsar significance diagrams for all metric ranks across all 6 data sets. The diagrams display the performance rank of each technique along with a line segment representing its corresponding critical difference (*CD* = 10.08).
A detailed description of the diagrammatic setup can be found in the previous chapter (cf. Chapter 4.5).

Rank	Technique	MAE	RMSE	AUC	AOC	R^2	r	ρ	τ	MR_{cal}	MR_{dis}	MR
1	LSSVM	7.5	3.5	3.3	3.5	3.5	3.7	3.3	4.1	4.5	3.6	4.1
2	ANN	3.2	2.8	5.0	2.5	2.7	3.1	7.0	7.1	2.8	5.5	4.2
3	OLS+LS-SVM	7.5	4.2	3.5	3.9	4.2	4.5	4.3	4.7	4.9	4.3	4.6
4	LOG+ANN	6.0	4.2	6.8	4.1	4.2	4.2	6.3	6.5	4.6	6.0	5.3
5	OLS+ANN	9.0	6.5	3.6	4.7	4.3	4.3	6.2	6.3	6.1	5.1	5.6
6	LOG+LS-SVM	7.5	6.4	4.6	6.4	6.5	5.2	5.2	4.9	6.7	5.0	5.8
7	OLS+RT	6.8	4.3	5.3	6.0	6.0	6.2	7.0	7.7	5.8	6.5	6.2
8	RT	8.6	7.0	12.9	7.4	7.0	7.8	7.3	4.7	7.5	8.2	7.8
9	LOG+RT	9.7	9.4	10.0	9.4	9.3	9.3	9.3	9.2	9.5	9.5	9.5
10	LOG+OLS	12.6	10.3	10.7	9.8	9.9	11.7	11.7	11.8	10.6	11.5	11.1
11	LOG+B-OLS	6.5	12.0	11.8	12.0	12.0	11.2	13.1	13.3	10.6	12.3	11.5
12	OLS	*13.9*	10.6	9.3	10.5	10.5	11.8	12.8	13.3	11.4	11.8	11.6
13	B-OLS	7.8	*14.3*	10.3	*14.8*	*14.0*	*13.8*	11.0	11.3	12.7	11.6	12.2
14	LOG+BC-OLS	8.2	*14.1*	13.0	*14.1*	*13.7*	13.0	11.8	11.8	12.5	12.4	12.5
15	BC-OLS	9.8	*15.7*	*13.8*	*15.6*	*15.3*	*14.7*	10.7	11.0	14.1	12.5	*13.3*
16	BR	*14.5*	*12.9*	*14.1*	*13.3*	*15.3*	*14.5*	11.8	11.5	14.0	13.0	*13.5*
17	LOG+BR	*13.9*	*14.9*	*14.9*	*15.0*	*14.6*	*14.2*	*14.2*	13.8	14.6	14.3	*14.4*

TABLE 5.10: Average rankings (AR) and meta-rankings (MR) across all metrics and data sets

Despite clear and consistent differences between regression techniques in terms of R^2, most techniques do not differ significantly according to the Nemenyi test. Nonetheless, failing to reject the null hypothesis that two techniques have equal performances does not guarantee that it is true. For example, Nemenyi's test is unable to reject the null hypothesis that ANN and OLS have equal performances although ANN consistently performs better than OLS. This can mean that the performance differences between these two are just due to chance. But the result could also be a Type II error. Possibly the Nemenyi test does not have sufficient power to detect a significant difference, given a significance level of $\alpha = 0.05$, 6 data sets and 17 techniques. The insufficient power of the test can be explained by the use of a large number of techniques in contrast with a relatively small number of data sets. (Normal probability plots for the OLS models across each of the data sets can be found in APPENDIX A5 at the end of this thesis).

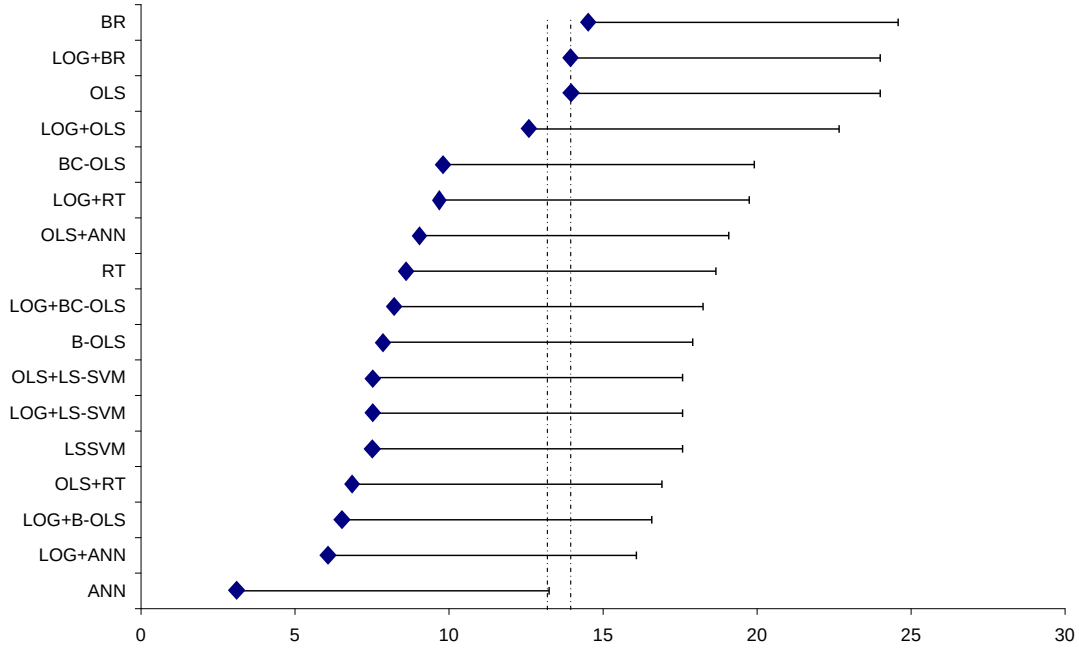


FIGURE 5.4: Demsar's significance diagram for MAE based ranks across 6 data sets

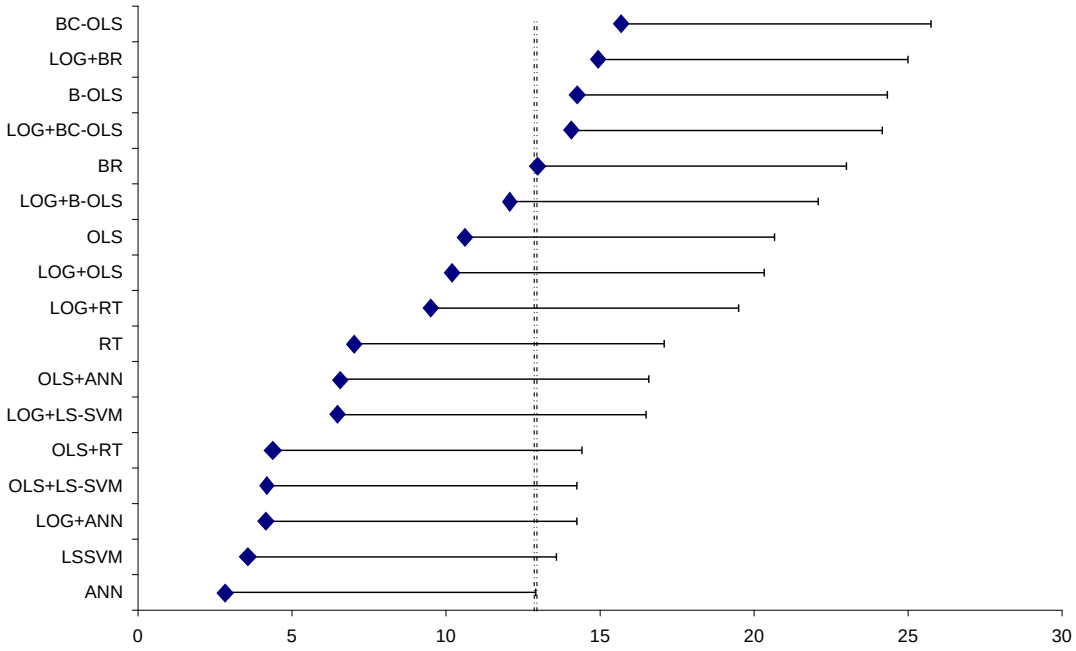


FIGURE 5.5: Demsar's significance diagram for RMSE based ranks across 6 data sets

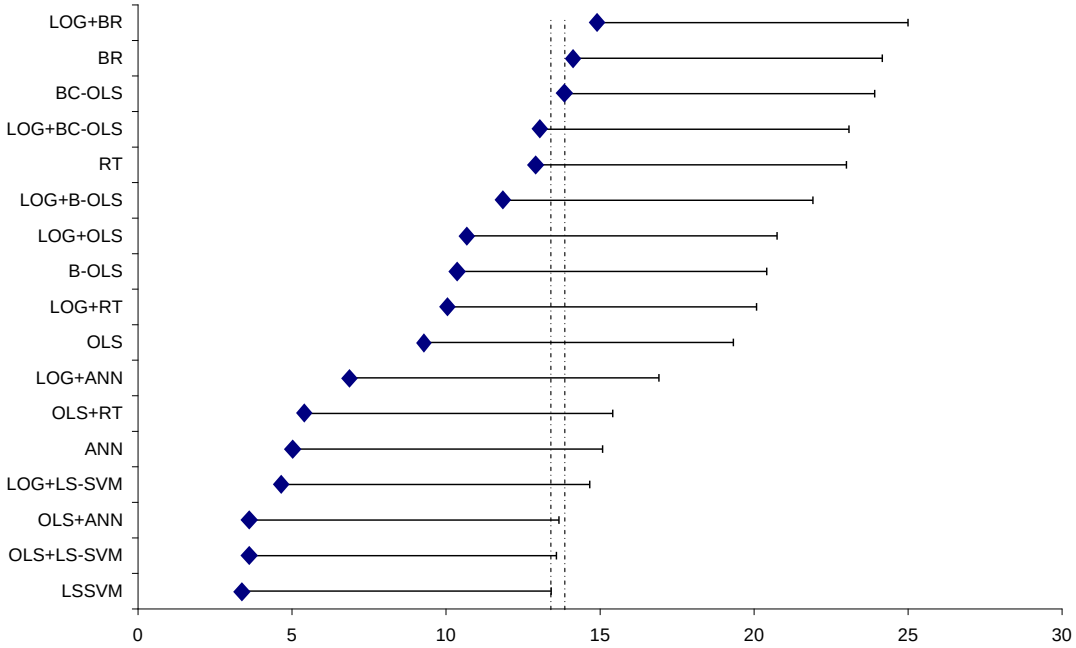


FIGURE 5.6: Demsar's significance diagram for AUC based ranks across 6 data sets

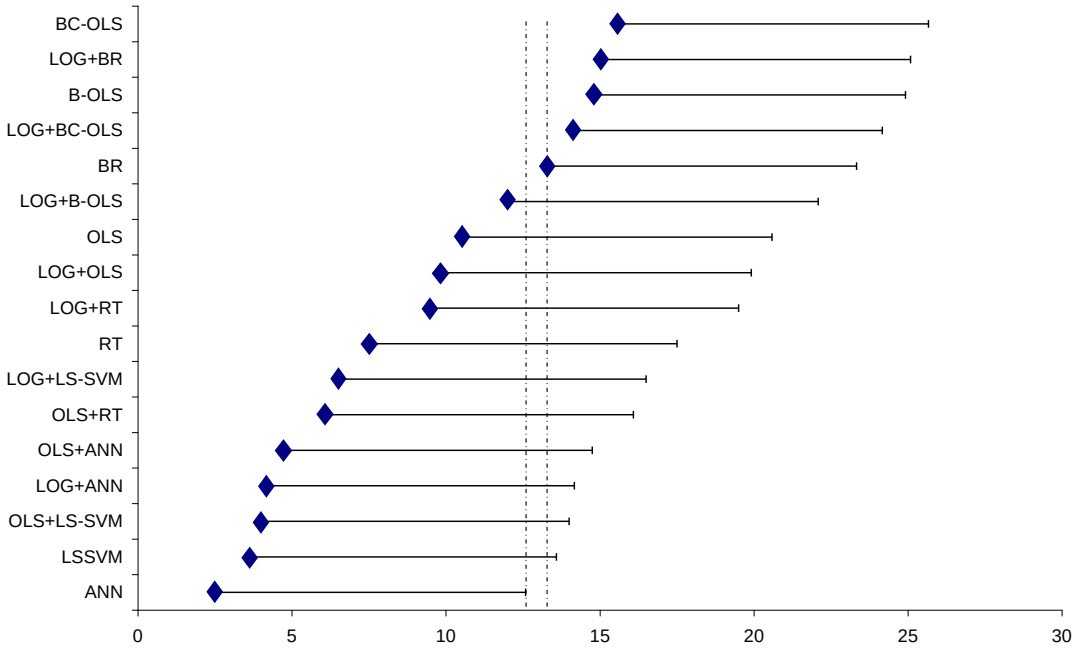


FIGURE 5.7: Demsar's significance diagram for AOC based ranks across 6 data sets

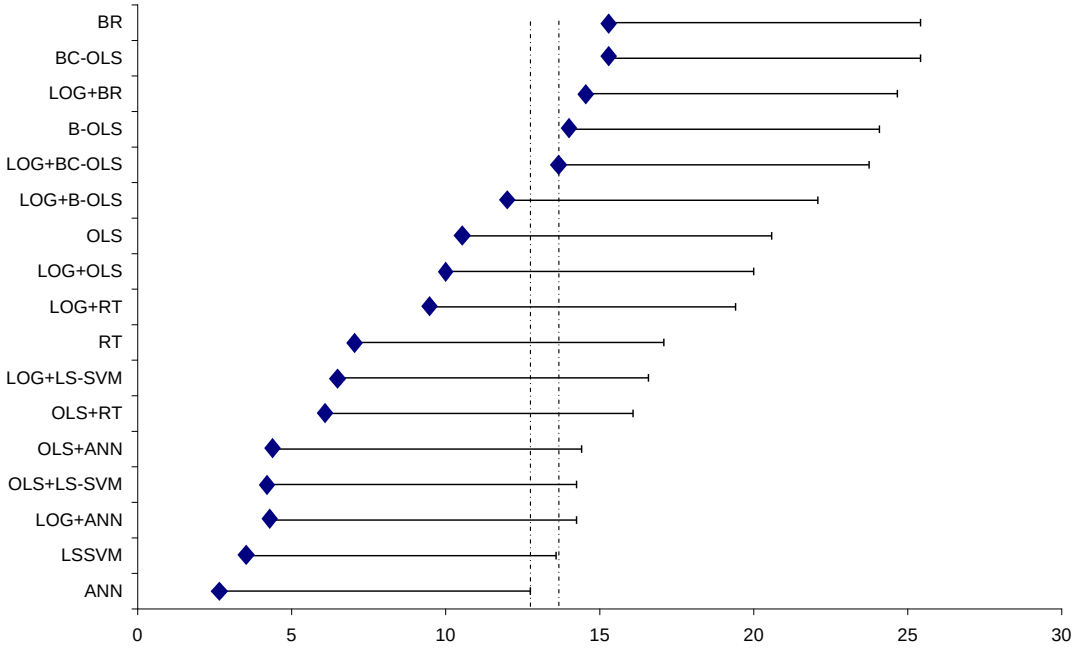


FIGURE 5.8: Demsar's significance diagram for R^2 based ranks across 6 data sets

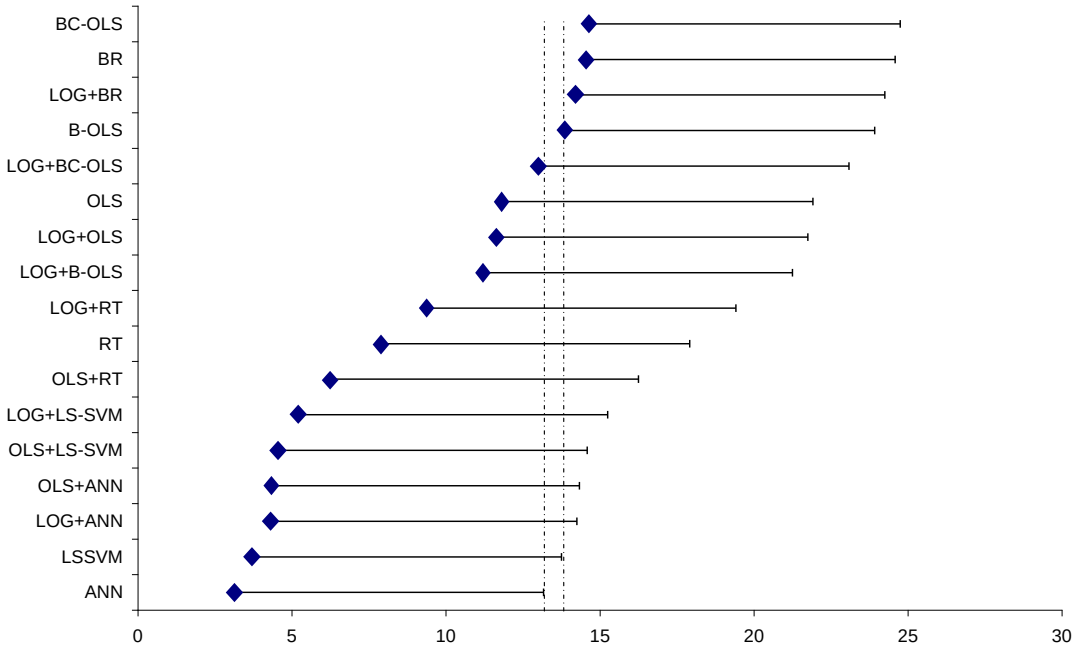


FIGURE 5.9: Demsar's significance diagram for r based ranks across 6 data sets

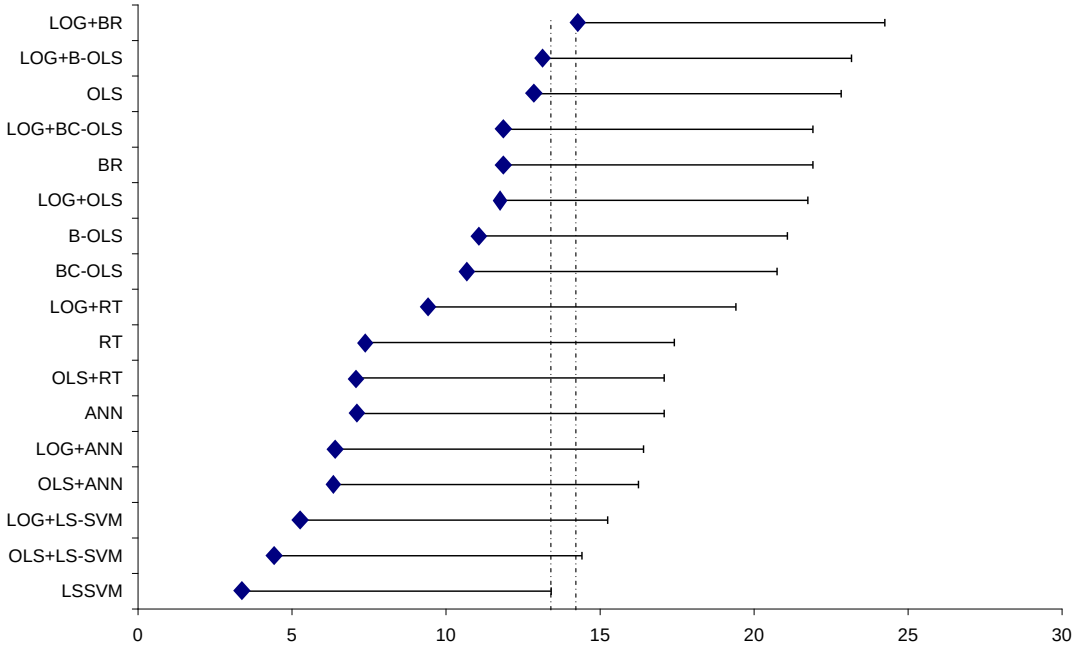


FIGURE 5.10: Demsar's significance diagram for ρ based ranks across 6 data sets

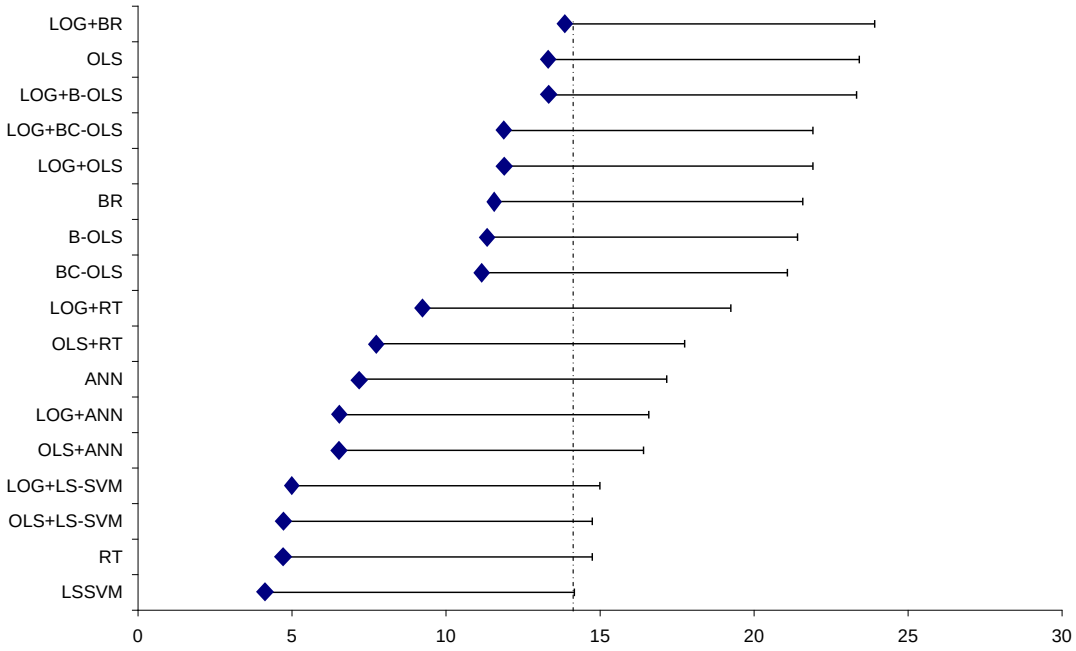


FIGURE 5.11: Demsar's significance diagram for τ based ranks across 6 data sets

5.6 Conclusions and recommendations for further work

This chapter evaluates the estimation of LGD through the use of 17 regression techniques on 6 real life retail lending data sets from major international banking institutions. The average predictive performance of the models in terms of R^2 ranges from 4 % to 43 %, which indicates that most resulting models do not have satisfactory explanatory power. Nonetheless, a clear trend can be seen that non-linear techniques such as artificial neural networks and support vector machines in particular give higher performances than more traditional linear techniques. This indicates the presence of non-linear interactions between the independent variables and the LGD, contrary to some studies in PD modelling (Baesens, et al. 2003) where the difference between linear and non-linear techniques is not that explicit. Given the fact that LGD has a bigger impact on the minimal capital requirements than PD, we demonstrated the potential and importance of applying non-linear techniques, preferably in a two-stage context to obtain comprehensibility as well, for LGD modelling. The findings presented in this chapter also go some way in agreeing with the findings presented in Qi and Zhao (2011), where it was shown that non-parametric techniques such as regression trees and neural networks gave improved model fit and predictive accuracy over parametric methods.

There is considerable evidence that the macro-economy affects the client's credit risk behaviour so it might be an interesting topic of further research to examine the influence of macro-economic variables (Bellotti & Crook, 2009), both in the context of improving LGD models as for stress testing. Finally, one could also try to add comprehensibility to well-performing black box models with rule extraction techniques to gain more insight (Martens, et al. 2007, Martens, et al. 2009).

Chapter 6

6 Regression Model Development for Credit Card Exposure At Default (EAD)

Under the Basel II requirements for the advanced internal ratings based approach (AIRB) banks must estimate and empirically validate their own models for Probability of Default (PD), Loss Given Default (LGD) and Exposure at Default (EAD). However, to date, the majority of academic literature has focused on the estimation and validation of PD and LGD models, with little work conducted on EAD modelling. In this chapter, we develop and compute a series of models for predicting Exposure At Default (EAD). For off-balance-sheet items, for example credit cards, to calculate the EAD one requires the committed but unused loan amount times a credit conversion factor (CCF). Ordinary least squares (OLS), binary logit and cumulative logit regression models, as well as an OLS with Beta transformation model, are estimated and compared with the main aim of finding the most robust and comprehensible model for the prediction of the CCF. Finally a direct estimation of EAD, using an OLS model, will be analysed.

A real-life data set with monthly balance amounts for clients over the period 2001-2004 will be used in the building and testing of the regression models. Parameter estimates and comparative statistics will be given to determine the best overall model. The findings from this study indicate that a marginal improvement in the coefficient of determination can be achieved with the use of a binary logit model over a traditional OLS model in the estimation of the CCF. It is also concluded that although the predictive power of the CCF is relatively weak across all of the models employed, when this predicted value is applied

to the EAD formulation to predict the actual EAD value, the predictive power is fairly strong. Interestingly the use of a direct estimation of EAD shows an increase in predictive power over first estimating a CCF and applying the CCF to the formulation.

6.1 Introduction

A detailed background and introduction to the topic of Exposure at Default (EAD) along with motivations for the work can be found in Chapter 1 of this thesis.

The purpose of this chapter will be to look at the estimation and validation of this credit conversion factor (CCF) in order to correctly estimate the off-balance sheet EAD. We also aim to gain a better understanding of the variables that drive the prediction of the CCF for consumer credit. To achieve this, predictive variables that have previously been suggested in the literature (Moral, 2006) will be constructed, along with a combination of new and potentially significant variables. We also aim to identify whether an improvement in predictive power can be achieved over ordinary least squares regression by the use of binary logit and cumulative logit regression models, as well as an OLS with Beta transformation model. The reason why we propose these two logit models is that recent studies (e.g. Jacobs, 2008) have shown that the CCF exhibits a bi-modal distribution with two peaks around 0 and 1, and a relatively flat distribution between those peaks. This non-normal distribution is therefore less suitable for modelling with traditional ordinary least squares (OLS) regression. The motivation for using an OLS with Beta transformation model is that it accounts for a range of distributions including a U-shaped distribution. We will also trial a direct OLS estimation of the EAD and use it as a comparison to estimating a CCF and applying it to the EAD formulation.
The purpose of this experimental setup is to extend the current literature and to better inform practitioners as to the potential techniques that can be applied in the estimation of CCF and the resulting EAD. It also aims to explore the practicalities of using OLS models for estimating the bi-modal distribution displayed by CCF and the potential of binning this distribution for the use of logistic and cumulative logistic regression models.

The remainder of this chapter is organised as follows. Section 6.2 outlines the proposed regression techniques that will be used in the estimation of the CCF. Section 6.3 details the empirical set up and data set used. Section 6.4 highlights the results of the regression

techniques in the estimation of the CCF. Finally, section 6.5 details the conclusions and recommendations that can be drawn from the results of the empirical study.

6.2 Overview of techniques

For the detailing of the techniques implemented in the estimation of the CCF value, the dependent variable y (i.e. the value of the CCF) for observation i is represented as y_i.

6.2.1 Ordinary Least Squares (OLS)

See Chapter 3 for a detailed overview of Ordinary Least Squares (OLS) regression.

6.2.2 Binary and Cumulative Logit models (LOGIT & CLOGIT)

The CCF distribution is often characterised by a peak around CCF = 0 and a further peak around CCF = 1 (cf. Infra, FIGURES 6.1 and 6.2). This non-normal distribution can lead to inaccurate linear regression models. Therefore, we propose the use of binary and cumulative logit models in an attempt to resolve this issue by grouping the observations for the CCF into two categories for the binary logit model and three categories for the cumulative logit model. For the binary response variable, two different splits will be tried: the first is made according to the mean of the CCF distribution ($\text{Class}\,0: CCF < \overline{CCF}$; $\text{Class}\,1: CCF \geq \overline{CCF}$) and the second is made based on whether the CCF is less than 1 ($\text{Class}\,0: CCF < 1$, $\text{Class}\,1: CCF \geq 1$). For the cumulative logit model, the CCF is split into three levels, i.e. $\text{Class}\,0: CCF = 0$, $\text{Class}\,1: 0 < CCF < 1$ and $\text{Class}\,2: CCF = 1$.

Binary logistic and cumulative logistic regression are derived in Chapter 3 of this thesis.

6.2.3 Ordinary Least Squares with Beta Transformation (B-OLS)

See Chapter 3 for a detailed overview of the Ordinary Least Squares with Beta Transformation (B-OLS) model.

6.3 Empirical set-up and data sets

The data set used was obtained from a major financial institution in the UK and contains monthly data on credit card usage for a three-year period (January 2001 – December 2004). Here, we define a default to have occurred on a credit card when a charge off has been made on that account (a charge off in this case is defined as the declaration by the creditor that an amount of debt is unlikely to be collected, declared at the point of 180 days or 6 months without payment). In order to calculate the CCF value, the original data set has been split into two twelve-month cohorts, with the first cohort running from November 2002 to October 2003 and the second cohort from November 2003 to October 2004. The cohort approach groups defaulted facilities into discrete calendar periods, in this case 12-month periods, according to the date of default. Information is then collected regarding risk factors and drawn/undrawn amounts at the beginning of the calendar period and drawn amount at the date of default. We have chosen the cohorts to begin in November and end in October as we wanted to reduce the effects of any seasonality on the calculation of the CCF.

The characteristics of the cohorts used in evaluating the performance of the regression models are given below in TABLE 6.1:

	Data set size (number of defaults)	Mean CCF (before winsorisation)	Standard Deviation (before winsorisation)	Mean CCF (after winsorisation)	Standard Deviation (after winsorisation)
COHORT1 (November 2002 – October 2003)	4,039	0.4055	2.7512	0.4901	0.4651
COHORT2 (November 2003 – October 2004)	6,232	0.5849	2.8124	0.5313	0.4626

TABLE 6.1: Characteristics of Cohorts for EAD data set

COHORT1 will be used to train the regression models, while COHORT2 will be used to test the performance of the model (out-of-time testing).
Both data sets contain variables detailing the type of defaulted credit card product and the following monthly variables: advised credit limit, current balance, the number of days delinquent and the behavioural score.

The following variables suggested in Moral (2006) were then computed based on the monthly data found in each of the cohorts, where t_d is the default date and t_r is the reference date (i.e. the start of the cohort):

- Committed amount, $L(t_r)$: the advised credit limit at the start of the cohort;
- Drawn amount, $E(t_r)$: the exposure at the start of the cohort;
- Undrawn amount, $L(t_r) - E(t_r)$: the advised limit minus the exposure at the start of cohort;
- Credit percentage usage, $\frac{E(t_r)}{L(t_r)}$: the exposure at the start of the cohort divided by the advised credit limit at the start of the cohort;
- Time to default, $(t_d - t_r)$: the default date minus the reference date (in months);
- Rating class, $R(t_r)$: the behavioural score at the start of the cohort, binned into four discrete categories 1: AAA-A; 2: BBB-B; 3: C; 4: UR (unrated).

The target variable was computed as follows:

- Credit conversion factor, CCF_i: calculated as the actual EAD minus the drawn amount at the start of the cohort divided by the advised credit limit at the start of the cohort minus the drawn amount at the start of the cohort, i.e. :

$$CCF_i = \frac{E(t_d) - E(t_r)}{L(t_r) - E(t_r)}. \qquad (6.1)$$

In addition to the aforementioned variables, we constructed a set of additional variables that could potentially increase the predictive power of the regression models implemented. These extra variables created are:

- Average number of days delinquent in the previous 3 months, 6 months, 9 months and 12 months. We expect the higher the number of days delinquent closer to default date, the higher the CCF value will be;
- Increase in committed amount: binary variable indicating whether there has been an increase in the committed amount since 12 months prior to the start of the cohort. We expect an increase in the committed amount to increase the value of the CCF;
- Undrawn percentage, $\frac{L(t_r) - E(t_r)}{L(t_r)}$: the undrawn amount at the start of the cohort divided by the advised credit limit at the start of the cohort. We expect higher ratios to result in a decrease in the value of the CCF;
- Absolute change in drawn, undrawn and committed amount: variable amount at t_r minus the variable amount 3 months, 6 months or 12 months prior to t_r;
- Relative change in drawn, undrawn and committed amount: variable amount at t_r minus the variable amount 3 months, 6 months or 12 months prior to t_r, divided by the variable amount 3 months, 6 months or 12 months prior to t_r, respectively.

The potential predictiveness of all the variables proposed in this chapter will be evaluated by calculating the information value (IV) based on their ability to separate the CCF value into either of two classes, $0: CCF < \overline{CCF}$ (non-event), and $1: CCF \geq \overline{CCF}$ (event). After binning input variables using an entropy-based procedure, implemented in SAS Enterprise Miner, the information value of a variable with k bins is given by:

$$IV = \sum_{i=1}^{k} \left[\left(\frac{n_1(i)}{N_1} - \frac{n_0(i)}{N_0} \right) \ln \left(\frac{n_1(i)/N_1}{n_0(i)/N_0} \right) \right], \tag{6.2}$$

where $n_0(i), n_1(i)$ denote the number of non-events and events in bin i, and N_0, N_1 are the total number of non-events and events in the data set, respectively.
This measure allows us to do a preliminary screening of the relative potential contribution of each variable in the prediction of the CCF.

The distribution of the raw CCF for the first Cohort (COHORT1) is shown below in FIGURE 6.1:

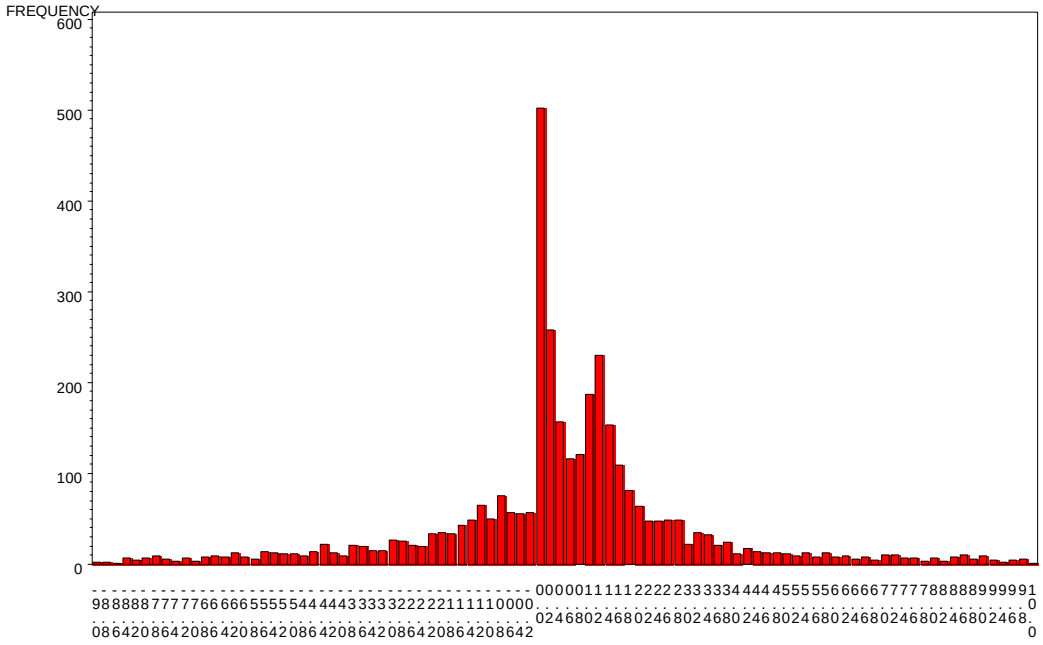


FIGURE 6.1: Raw CCF distribution (x-axis displays a snapshot of the CCF values from the period of -9 to 10)

The raw CCF displays a substantial peak around 0 and a slight peak at 1 with substantial tails either side of these points. (FIGURE 6.1 displays a snapshot of CCF values in the period -9 to 10. This snapshot boundary has been selected to allow for the visualisation of the CCF distribution.) Values of $CCF > 1$ can occur when the actual EAD is greater than the advised credit limit, whereas values of $CCF < 0$ can occur when both the drawn amount and the EAD exceed the advised credit limit or where the EAD is smaller than the drawn amount. In practice this occurs as the advised credit limit and drawn amount are measured at a time period, (t_r), prior to default and therefore at (t_d) the advised credit limit maybe higher or lower than at (t_r). Extremely large positive and negative

values of CCF can also occur if the drawn amount is slightly above or below the advised credit limit, e.g.:

$$CCF_i = \frac{E(t_d) - E(t_r)}{L(t_r) - E(t_r)} = \frac{3300 - 3099.9}{3100 - 3099.9} = 2001 \tag{6.3}$$

$$CCF_i = \frac{E(t_d) - E(t_r)}{L(t_r) - E(t_r)} = \frac{3000 - 3500}{4000 - 3500} = -1 \tag{6.4}$$

As in Jacobs, (2008) and Qi, (2009) it therefore seems reasonable to winsorise the data so that the CCF can only fall between values of 0 and 1. FIGURE 6.2 displays the same CCF value winsorised at 0 and 1:

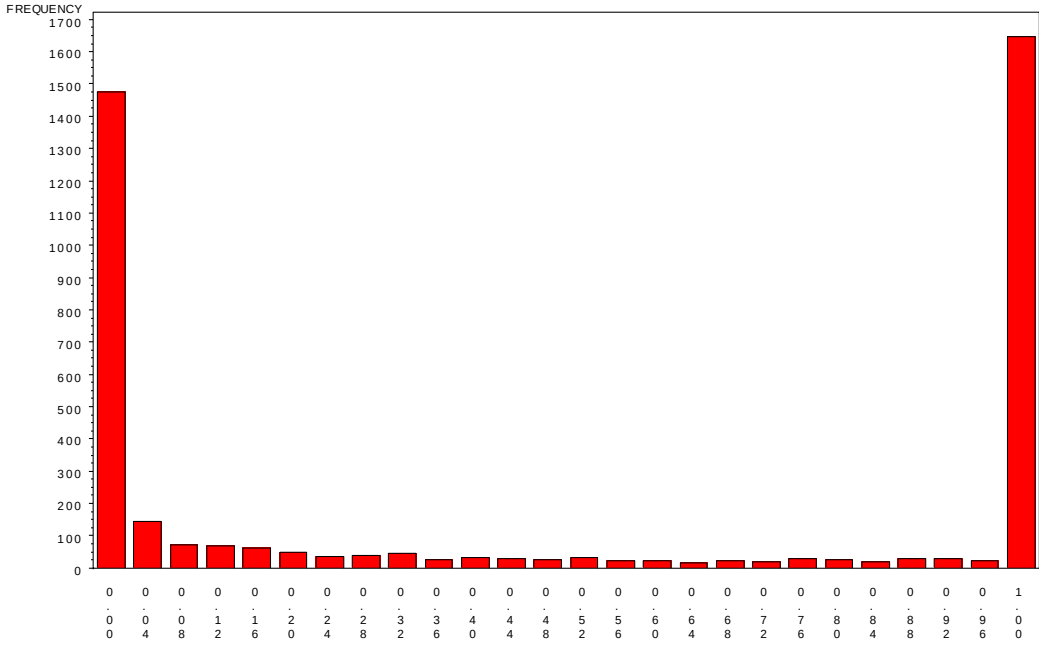


FIGURE 6.2: CCF distribution winsorised (between 0 and 1)

The winsorised CCF (FIGURE 6.2) yields a bimodal distribution with peaks at 0 and 1, and a relatively flat distribution between the two peaks. This bears a strong resemblance to the distributions identified in loss given default modelling (LGD) (Thomas *et al*, 2010). In our estimation of the CCF we will be using this limited CCF between 0 and 1, similarly to Jacobs (2008).

The OLS, B-OLS, LOGIT and CLOGIT models were estimated using SAS. Each model was built on the first Cohort data set (COHORT1) and then tested on the second Cohort data set (COHORT2).

A stepwise variable selection method was used in the construction of all three regression models with the aim of selecting only the most predictive input variables for the estimation of the CCF. The threshold level for the variables to enter and remain in the model using the stepwise procedure was a p-value of 0.01. For the LOGIT and CLOGIT models the resulting predicted probabilities were taken as the values for the CCF.

The following performance metrics were used to compare the regression techniques:

6.3.1 Coefficient of Determination (R^2)

The coefficient of determination (R^2) (Draper and Smith, 1998) can be defined as 1 minus the fraction of unexplained variance, i.e.:

$$R^2 = 1 - \frac{SS_{err}}{SS_{tot}}, \tag{6.5}$$

where $SS_{err} = \sum_{i=1}^{l} \left(y_i - f\left(\mathbf{x}_i\right) \right)^2$, $SS_{tot} = \sum_{i=1}^{l} \left(y_i - \bar{y} \right)^2$, and $\bar{y}$ is the mean of the observed CCF value. Although R^2 is usually a number from 0 to 1, R^2 could also yield negative values when the model prediction is worse than using the mean $\bar{y}$ from the training set as a prediction.

In order to calculate the performance metrics on the categorical predictions made by the LOGIT and CLOGIT models, first a continuous prediction value must be obtained. This is achieved by multiplying the probability of being in each of the bins by the average CCF value for each of those respective bins and summing the result, thus obtaining an expected value of CCF. After this value has been computed, the resulting value is then used in the calculation of the performance metrics.

6.3.2 Pearson's Correlation Coefficient (r)

Pearson's correlation coefficient (see e.g. Cohen et al, 2002) is defined as the sum of the products of the standard scores of the observed and predicted values divided by the degrees of freedom.

6.3.3 Spearman's Correlation Coefficient (ρ)

Spearman's ρ (see e.g. Cohen et al, 2002) is defined as the Pearson's r applied to the rankings of predicted and observed values.

6.3.4 Root Mean Squared Error (RMSE)

The Root Mean Squared Error (RMSE) is defined as the square root of the average of the squared difference between predictions and obtained values:

$$RMSE = \sqrt{\frac{1}{l}\sum_{i=1}^{l}\left(f\left(\mathbf{x}_i\right) - y_i\right)^2} \tag{6.6}$$

6.4 Results and discussion

In this section we will begin by analysing the input variables and their relationship to the dichotomised CCF value ($0: CCF < \overline{CCF}$; $1: CCF \geq \overline{CCF}$). The following table displays the resulting information value for each variable, ranked from most to least predictive:

Variable	Information Value
Credit percentage usage	1.825
Undrawn percentage	1.825
Undrawn	1.581
Relative change in undrawn amount (12 months)	0.696
Relative change in undrawn amount (6 months)	0.425
Relative change in undrawn amount (3 months)	0.343
Rating Class	0.233
Time-to-Default	0.226
Drawn	0.181
Absolute change in drawn amount (3 months)	0.114
Absolute change in undrawn amount (3 months)	0.089
Absolute change in undrawn amount (12 months)	0.083
Absolute change in undrawn amount (6 months)	0.072
Absolute change in drawn amount (6 months)	0.063
Relative change in drawn amount (3 months)	0.058
Absolute change in drawn amount (12 months)	0.054
Relative change in drawn amount (6 months)	0.049
Average number of days delinquent (9 months)	0.041
Average number of days delinquent (3 months)	0.040
Average number of days delinquent (6 months)	0.040
Relative change in drawn amount (12 months)	0.040
Average number of days delinquent (12 months)	0.032
Relative change in committed amount (12 months)	0.026

Absolute change in committed amount (12 months)	0.023
Absolute change in committed amount (3 months)	0.023
Relative change in committed amount (3 months)	0.022
Relative change in committed amount (6 months)	0.021
Absolute change in committed amount (6 months)	0.021
Increase in committed amount	0.018
Committed amount	0.017

TABLE 6.2: Information Values of constructed variables

Typically, input variables which display an information value greater than 0.1 are deemed to have a significant contribution in the prediction of the target variable. From this analysis, we can see that the majority of the relative and absolute changes in drawn, undrawn and committed amounts do not possess the same ability to discriminate between low and high CCFs as the original variable measures at reference time only. It is also clear from the results that the undrawn amount could be an important variable in the discrimination of the CCF value. It must be taken into consideration however that although the variables may display a good ability to discriminate between the low and high CCFs, the variables themselves are highly correlated with each other (see Table A6 in the APPENDIX).

Subsequently, we examine the performance of the models themselves in the prediction of the CCF. The following table (TABLE 6.3) reports the parameter estimates and p-values for the variables used by each of the regression techniques implemented. The parameter signs found in Jacobs, (2008) are also shown for comparative purposes. The five regression models detailed are: an OLS model implementing only the suggested predictive variables in Moral, (2006); an OLS model incorporating the additional variables after stepwise selection; an OLS with Beta transformation model; a binary logit model and a cumulative logit model. For the binary logit model the best class split found was to select $0: CCF < 1$ and $1: CCF \geq 1$. It is however important to note that little difference was found between the choices of class split for the binary model.

From TABLE 6.3, we can see that the best performing regression algorithm for all three performance measures is the binary logit model with an R^2 value of 0.1028. Although this R^2 value is low, it is comparable to the range of performance results previously reported in other work on LGD modelling (cf. Chapter 5). This result also re-affirms the proposed usefulness of a logit model for estimating CCFs in Valvonis (2008). It can also be seen that all five models are quite similar in terms of variable significance levels and positive/negative signs. There does however seem to be some discrepancy for the Rating class variable, where the medium-range behavioural score band appears to be associated with the highest CCF's.

Variables	Coefficient sign reported in Jacobs, (2008)	OLS model (using only suggested variables in Moral, (2006))		OLS model (OLS) (additional variables)		OLS with Beta transformation (B-OLS)		Binary logit model (LOGIT)		Cumulative logit model (CLOGIT)	
		Parameter Estimate	P-value	Parameter Estimate	P-value	Parameter Estimate	P-value	Parameter Estimate	P-value	Parameter Estimate	P-value
Intercept 1		0.1830	<.0001	0.1365	<.001	-0.5573	<.0001	-1.5701	<.0001	0.6493	<.0001
Intercept 2										-0.5491	<.001
Credit percentage usage	–	-0.1220	<.001	-0.1260	<.001			-0.5737	<.001	-1.3220	<.0001
Committed amount	+	1.73E-05	<.0001	1.76E-05	<.0001	2.2E-05	<.0001	9.0E-05	<.0001	8.8E-05	<.0001
Undrawn	+	-8.68E-05	<.0001	-8.88E-05	<.0001	-1.1E-04	<.0001	-4.7E-04	<.0001	-3.6E-04	<.0001
Time-to-Default	+	0.0334	<.0001	0.0326	<.0001	0.0358	<.0001	0.1538	<.0001	0.1009	<.0001
Rating class	–										
Rating 1 (AAA-A) vs. 4 (UR)		0.1735	<.0001	0.2304	<.0001	0.2223	<.0001	0.4000	0.0069	-0.0772	0.5472
Rating 2 (BBB-B) vs. 4 (UR)		0.2483	<.0001	0.2977	<.0001	0.3894	<.0001	0.5885	<.0001	0.6922	<.0001
Rating 3 (C) vs. 4 (UR)		0.0944	<.0001	0.1201	<.0001	0.1664	<.0001	-0.2121	0.0043	-0.0157	0.8098
Average number of days delinquent in the last 6 months	N/A			0.0048	<.0001	0.0062	<.0001	0.0216	<.0001	0.0218	<.0001
Undrawn percentage	N/A					0.2784	<.0001				
Coefficient of Determination (R^2)		0.0982		0.0960		-0.0830		0.1028		0.0822	
Pearson's Correlation Coefficient (r)		0.3170		0.3144		0.3125		0.3244		0.2897	
Spearman's Correlation Coefficient (ρ)		0.2932		0.2943		0.3000		0.3283		0.2943	
Root Mean Squared Error (RMSE)		0.4393		0.4398		0.4833		0.4704		0.4432	

TABLE 6.3: Parameter estimates and P-values for CCF estimation on the COHORT2 data set

Of the additional variables we tested (e.g. absolute or relative change in the drawn amount, credit limit and undrawn amount), only 'Average number of days delinquent in the last 6 months' and 'Undrawn percentage' were retained by the stepwise selection procedures. This is most likely due to the fact that their relation to the CCF is already largely accounted for by the base model variables. Further to this, Table A6 in the APPENDIX details a correlation matrix for the inputs, indicating that for example the Drawn amount has a high positive correlation with the Committed Amount (0.782). It is also of interest to note that although one additional variable is selected in the stepwise procedure for the second OLS model, there is no increase in predictive power over the original OLS model.

A direct estimation of the un-winsorised CCF with the use of an OLS model was also undertaken. The results from this experimentation indicate that it is even harder to predict the un-winsorised CCF than the CCF winsorised between 0 and 1 with a predictive performance far weaker than the winsorised model. (When these results are applied to the estimation of the actual EAD an inferior result is also achieved).

With the predicted values for the CCF obtained from the five models, it is then possible to estimate the actual EAD value for each observation i in the COHORT2 data set, as follows:

$$EAD_i = E(t_r) + CCF_i.\left(L(t_r) - E(t_r)\right). \qquad (6.7)$$

This gives us an estimated "monetary EAD" value which can be compared to the actual EAD value found in the data set. For comparison purposes, a conservative estimate for the EAD $\left(\text{assuming } CCF = 1\right)$ is also calculated, as well as an estimate for EAD where the mean of the CCF in the first cohort is used (TABLE 6.4). The following table (TABLE 6.5) displays the predictive performance of this estimated EAD amount against the actual EAD amount:

Variables	Conservative estimate of EAD (CCF=1)	Estimate of EAD where CCF equals the mean CCF in first cohort
Coefficient of Determination (R^2)	0.5178	0.6486
Pearson's Correlation Coefficient (r)	0.7588	0.8062
Spearman's Correlation Coefficient (ρ)	0.6867	0.7354

TABLE 6.4: EAD estimates based on conservative and mean estimate for CCF

Variables	OLS model (using only previously suggested variables)	OLS model (including average number of days delinquent in the last 6 months)	OLS with Beta transformation (B-OLS)	Binary logit model (LOGIT)	Cumulative logit model (CLOGIT)
Coefficient of Determination (R^2)	0.6450	0.6431	0.8365	0.6344	0.6498
Pearson's Correlation Coefficient (r)	0.8049	0.8038	0.8000	0.8016	0.8068
Spearman's Correlation Coefficient (ρ)	0.7421	0.7405	0.7270	0.7387	0.7381

TABLE 6.5: EAD estimates based on CCF predictions against actual EAD amounts

It is quite clear from these results that although the predicted CCF value gave a relatively weak performance, when this value is applied to the calculation of the estimated EAD formulation a significant improvement over the conservative model can be made. It can also be noted that the application of the OLS with Beta transformation model gives a significantly higher value for the coefficient of determination (0.8365), although the correlation values are comparative to the other models. A possible reason for this is that even though the CCF has been winsorised prior to estimation, the B-OLS model's predictions are much closer to the real CCF values before winsorisation. Thus the B-OLS model produces a better actual estimate of the EAD. However, by simply applying the mean of the CCF, a similar result to the other predicted models can be achieved.

The direct estimation of the EAD, through the use of an OLS model, has also been taken into consideration, without the first estimation and application of a CCF. The results from this direct estimation of EAD are shown in TABLE 6.6, with the distribution for the direct estimation of EAD given in FIGURE 6.3: (The legend for FIGURES 6.3-6.7

details the frequency of values along the y-axis and the estimated EAD value along the x-axis)

Variables	OLS model (direct estimation of EAD)
Coefficient of Determination (R^2)	0.6608
Pearson's Correlation Coefficient (r)	0.8130
Spearman's Correlation Coefficient (ρ)	0.7493

TABLE 6.6: Direct Estimation of EAD

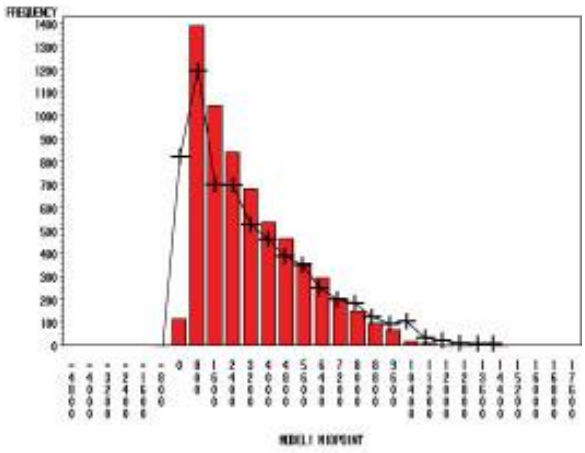


FIGURE 6.3: Distribution of direct estimation of EAD (the actual EAD amount present is indicated by the overlaid black line)

It is self-evident from the performance metrics and the produced distribution that a direct estimation of EAD without firstly estimating and applying a CCF can indeed produce reasonable estimations for the actual EAD. This goes someway in ratifying the findings show by Taplin et al (2007).

The following figures (FIGURES 6.4-6.7) display the distribution for the actual EAD amount present in COHORT2 and the estimated EAD values for the regression models. It is apparent from the predicted distributions that all five models approximate the actual EAD distribution very well. All three models do however somewhat underestimate the

number of observations at both ends of the distribution, corresponding to an EAD value of zero and values of EAD greater than 10,400. This is further evidence that although the regression models struggle to predict the actual CCF value, when this factor is applied to the EAD calculation a relatively good correlation can be achieved.

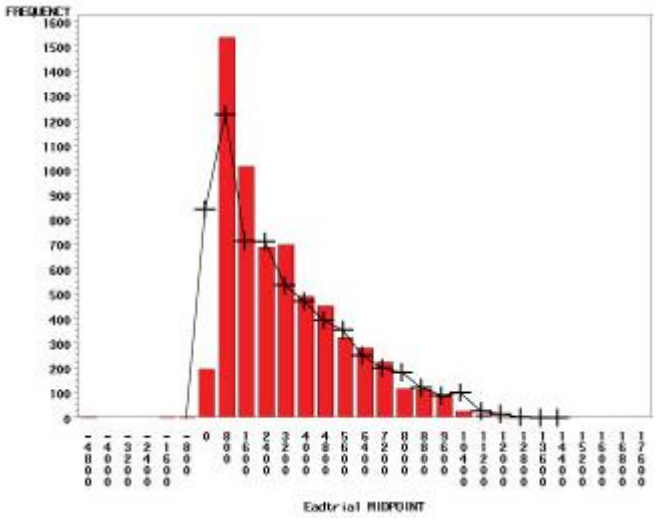


FIGURE 6.4: OLS base model predicted Exposure at Default (EAD) distribution (the actual EAD amount present is indicated by the overlaid black line)

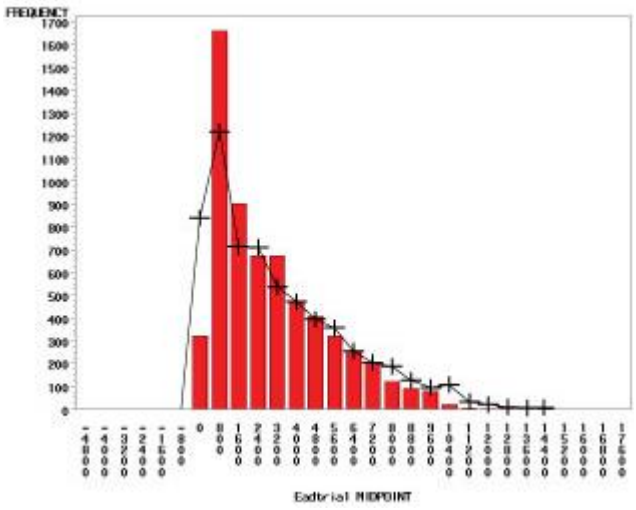


FIGURE 6.5: Binary LOGIT model predicted Exposure at Default (EAD) distribution (the actual EAD amount present is indicated by the overlaid black line)

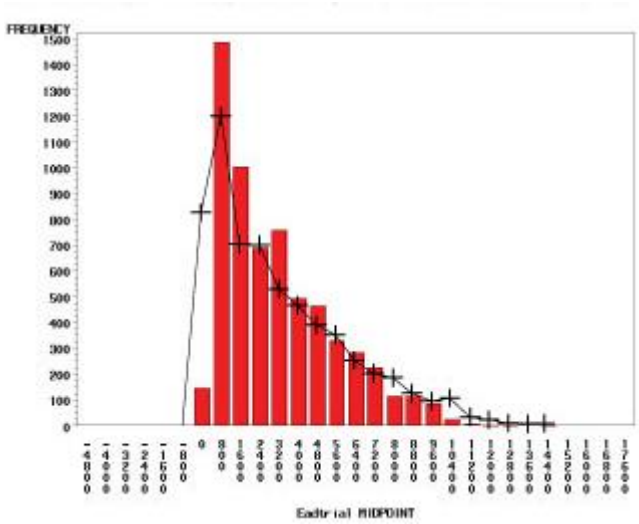


FIGURE 6.6: Cumulative LOGIT model predicted Exposure at Default (EAD) distribution (the actual EAD amount present is indicated by the overlaid black line)

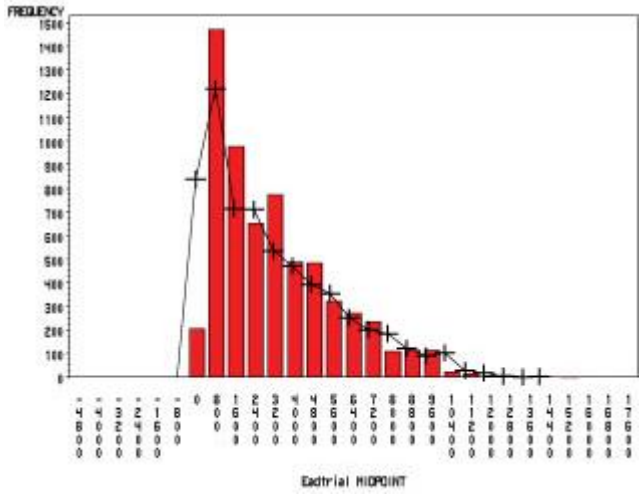


FIGURE 6.7: OLS with Beta Transformation model predicted Exposure at Default (EAD) distribution (the actual EAD amount present is indicated by the overlaid black line)

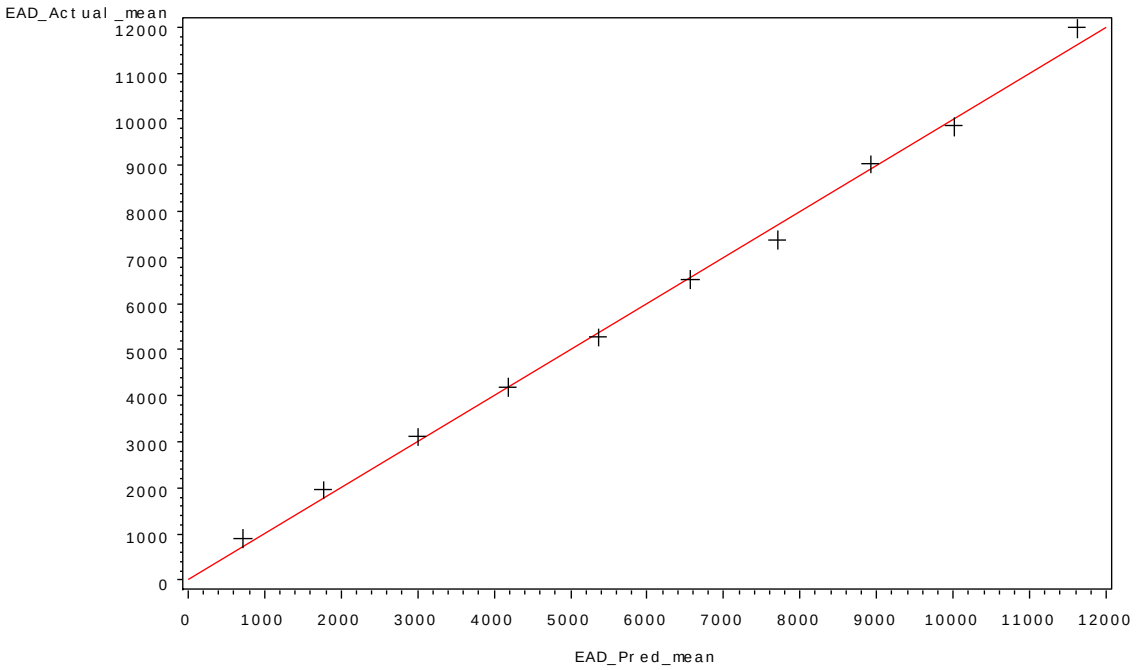


FIGURE 6.8: OLS base model plot for the Actual Mean EAD against Predicted Mean EAD across ten bins (R^2=0.9968)

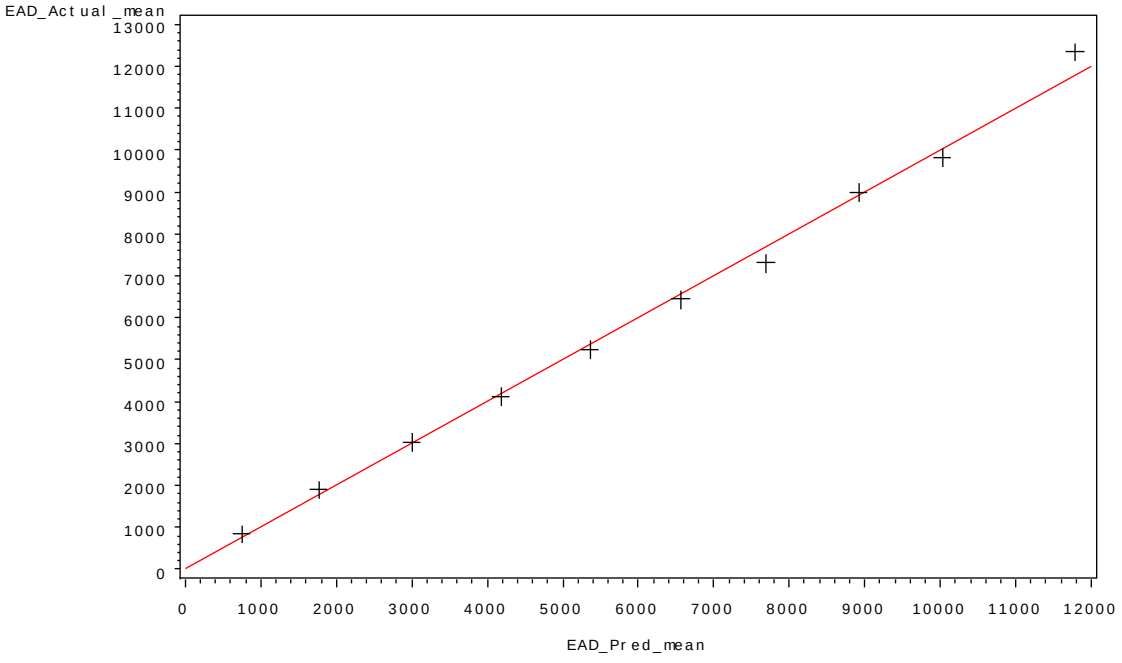


FIGURE 6.9: Binary LOGIT model plot for the Actual Mean EAD against the Predicted Mean EAD across ten bins (R^2=0.9944)

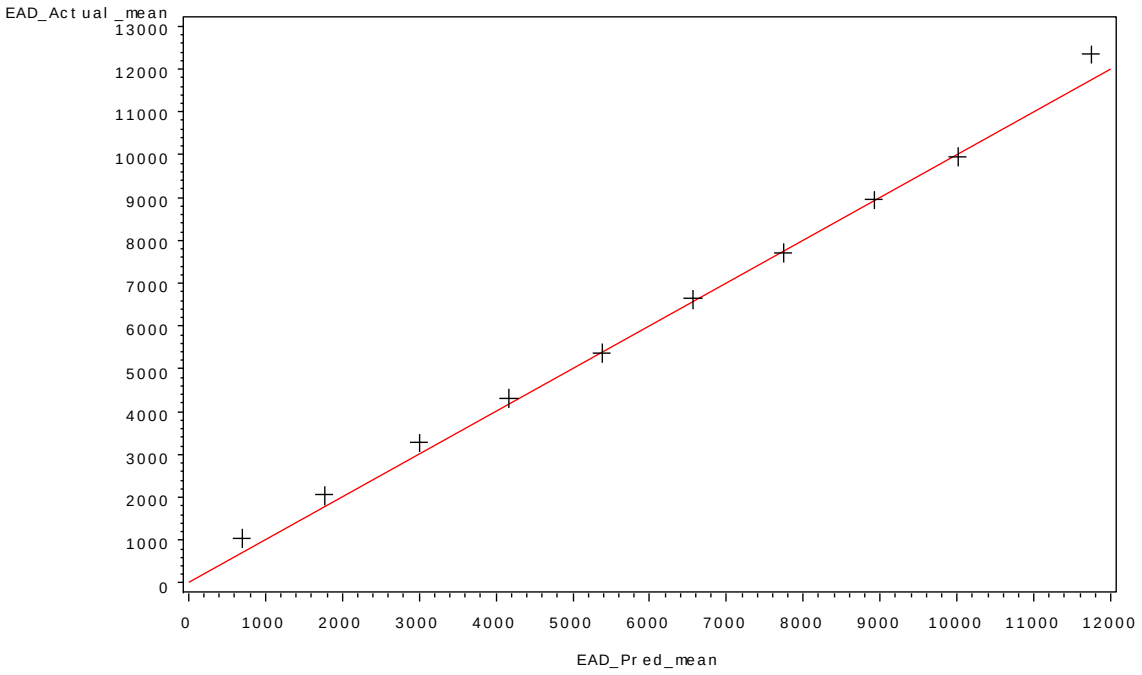


FIGURE 6.10: Cumulative LOGIT model plot for the Actual Mean EAD against the Predicted Mean EAD across ten bins (R^2=0.9954)

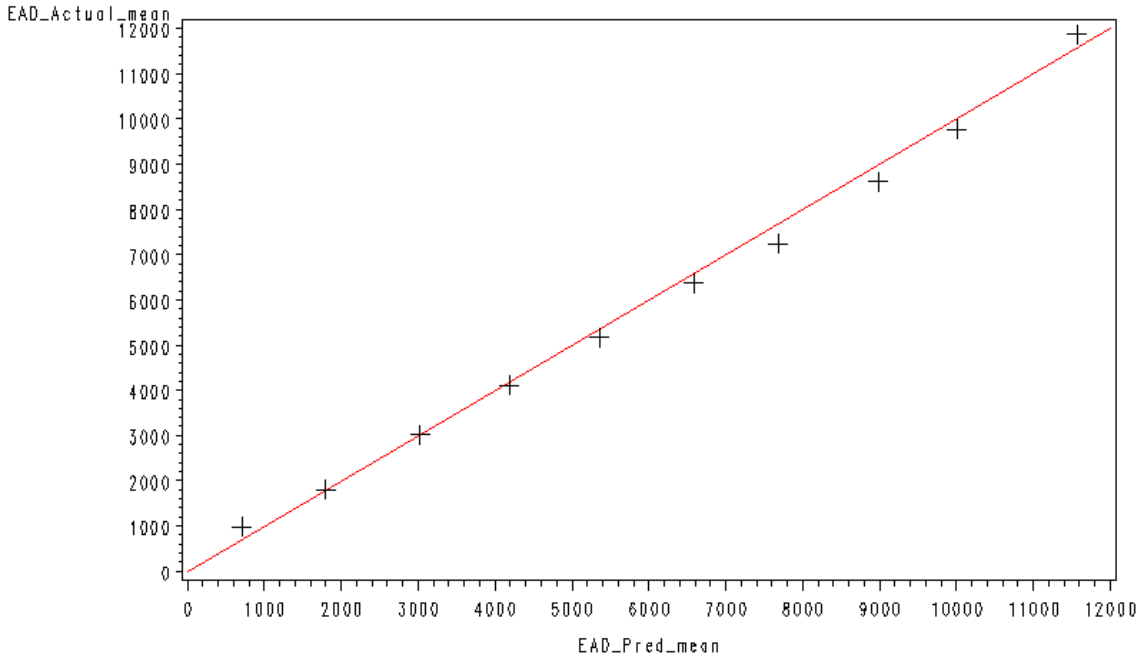


FIGURE 6.11: OLS with Beta Transformation model plot for the Actual Mean EAD against the Predicted Mean EAD across ten bins (R^2=0.9957)

FIGURES 6.8-6.11 display plots for the actual mean of the EAD against the predicted mean of the EAD across ten bins. (The legend for FIGURES 6.8-6.11 details the mean actual EAD along the y-axis and the mean predicted EAD along the x-axis across the 10 bins). The bins are created by splitting the distribution of the predicted EAD into ten bins of equal size. The plots show that the means for the actual and predicted EAD in bins one to ten are close to the diagonal for all three models, indicating that the predictions for the EAD well approximate actual EAD. The points that deviate slightly from the diagonal again occur at the left and right ends of the EAD.

Similarly to FIGUREs 6.8-6.11, FIGURES 6.12-6.15 display plots for the actual mean of the CCF against the predicted mean of the CCF across ten bins. (The legend for FIGURES 6.12-6.15 details the mean actual CCF along the y-axis and the mean predicted CCF along the x-axis across the 10 bins). From these plots it is clear that the regression models struggle to closely predict the values for CCF.

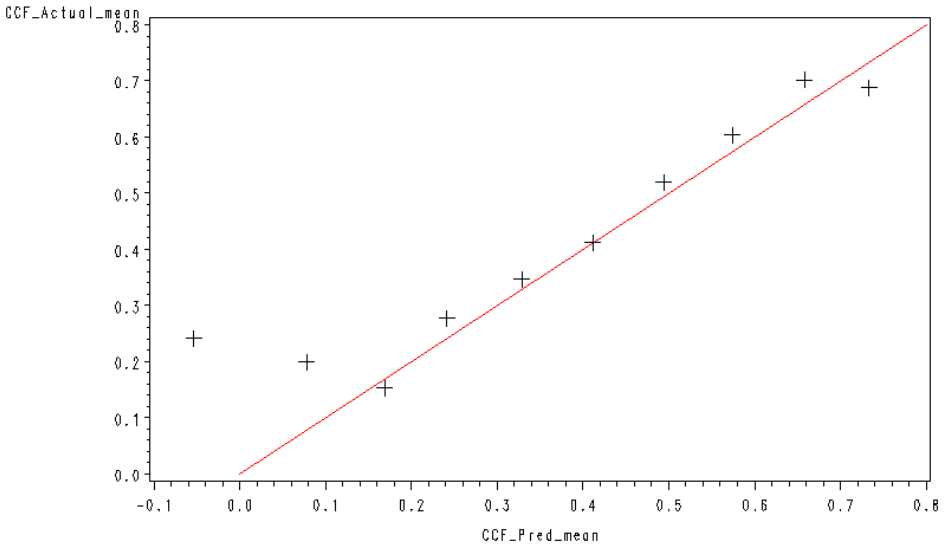


FIGURE 6.12: OLS base model plot for the Actual Mean CCF against the Predicted Mean CCF across ten bins (R^2=0.7061)

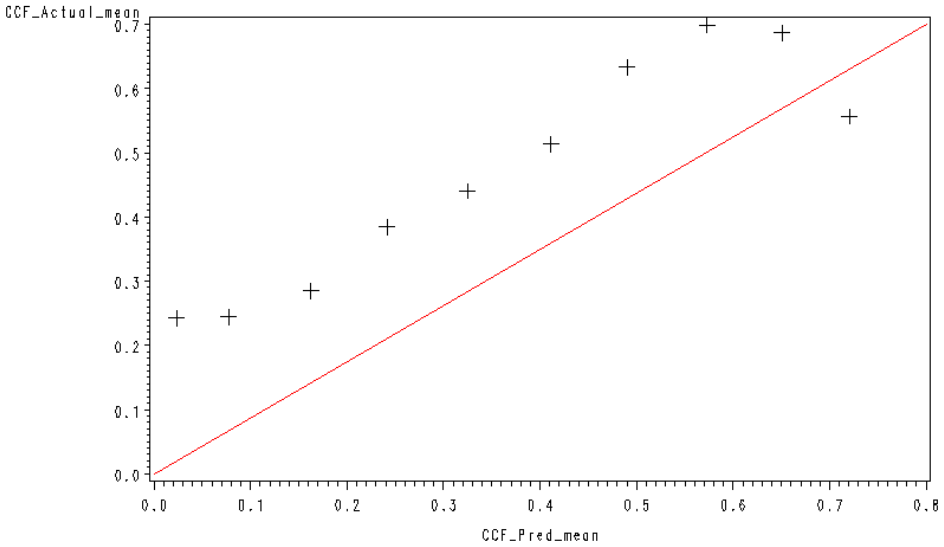


FIGURE 6.13: Binary LOGIT model plot for the Actual Mean CCF against the Predicted Mean CCF across ten bins (R^2=0.2867)

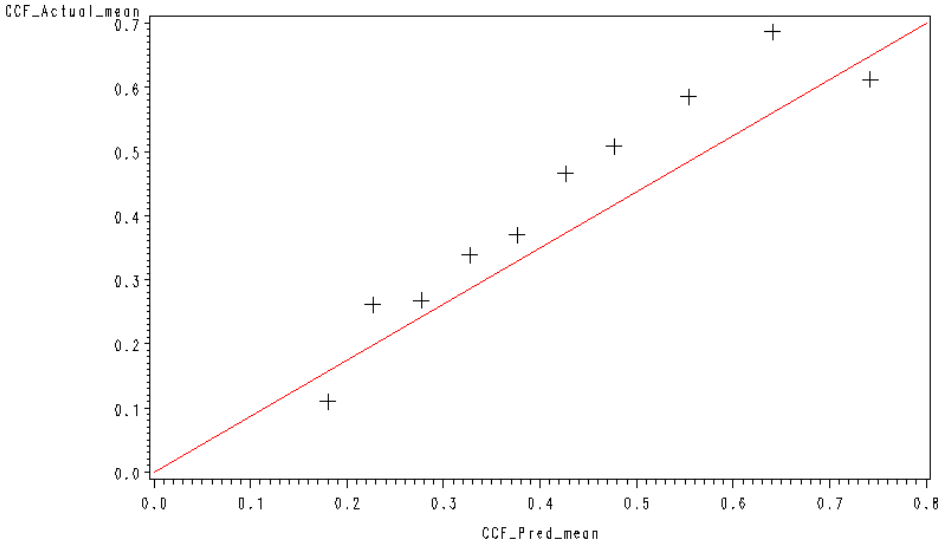


FIGURE 6.14: Cumulative LOGIT base model plot for the Actual Mean CCF against the Predicted Mean CCF across ten bins (R^2=0.9063)

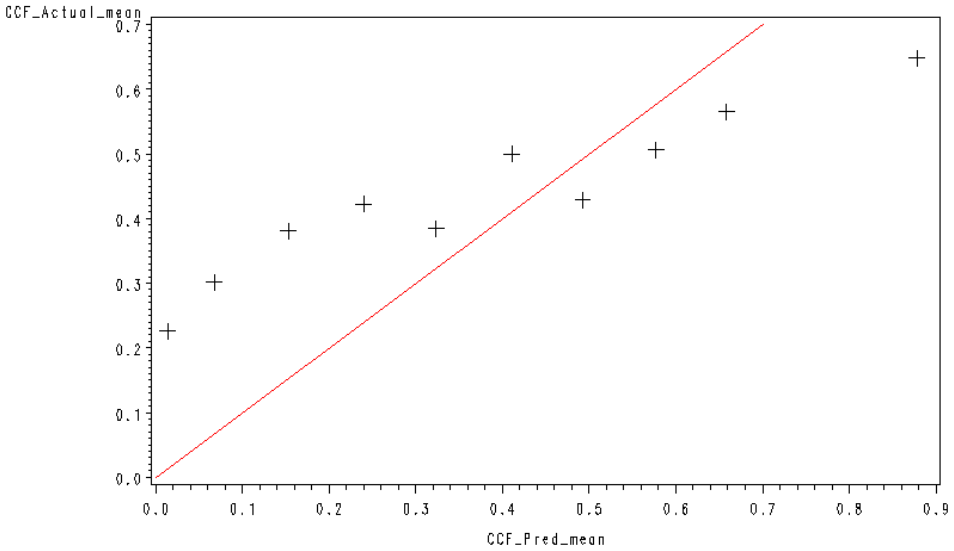


FIGURE 6.15: OLS with Beta Transformation model plot for the Actual Mean CCF against the Predicted Mean CCF across ten bins (R^2=0.9154)

6.5 Conclusions and recommendations for further work

In summary, this chapter has set out to develop comprehensible and robust regression models for the estimation of Exposure at Default (EAD) for consumer credit through the prediction of the credit conversion factor (CCF). An in-depth analysis of the predictive variables used in the modelling of the CCF has also been given, showing that previously acknowledged variables are significant and identifying a series of additional variables.

As the results show, a marginal improvement in the coefficient of determination can be achieved with the use of a binary logit model over a traditional OLS model. Interestingly the use of a cumulative logit model performs worse than both the binary logit and OLS models. The probable cause of this are the size of the peaks around 0 and 1 compared to the number of observations found in the interval between the two peaks. This therefore allows for more error in the prediction of the CCF via a cumulative three-class model.

Another interesting finding is that although the predictive power of the CCF is weak, when this predicted value is applied to the EAD formulation to predict the actual EAD value, the predictive power is fairly strong. In particular when the predictive values obtained through the application of the OLS with Beta transformation model were applied to the EAD formulation an improvement in the coefficient of determination was seen. Nonetheless, similar performance, in terms of correlations, could be achieved by a simple model that takes the average CCF of the previous cohort, showing that much of the explanatory power of EAD modelling derives from the current exposure.

With regards to the additional variables proposed in the prediction of the CCF only one, i.e. average number of days delinquent in the last 6 months, gave an adequate p-value, whilst undrawn percentage, potentially an alternative to credit percentage, was significant for the OLS with Beta transformation model. Even though the relative changes in the undrawn amount give reasonable information value scores, these variables do not prove to be significant in the regression models, probably due to their high correlation with the undrawn variable. This shows that the actual values at the start of the cohort already give a significant representation of previous activity in order to predict the CCF.

The contributions to the literature therefore are a greater understanding of the practicalities of applying OLS (with transformation) based models in the estimation of CCF, and the ability of these models compared to bucketing the CCF and using a logistic or cumulative logistic model. This chapter also goes someone to answering the question as to whether a direct estimation of EAD is appropriate instead of first estimating CCF and then using those values for calculating EAD. It is evident from the results presented here that a direct estimation of EAD without firstly estimating and applying a CCF can indeed produce reasonable estimations for the actual EAD. With regards to the issues highlighted in the literature review section of this thesis reinforces the findings by Taplin et al (2007) that a direct estimation of EAD could feasibly be more appropriate than first estimating CCF. The findings from this study not only agree with the findings by Moral (2006) but also contributes a new potentially significant variable in the estimation of CCF, which is average number of days delinquent in the last 6 months.

There is an obvious need for further research into the prediction of the exposure at default (EAD) value as this chapter can only go so far in its estimations. A more extensive study with multiple data sets over a longer timescale would be able to give more reliable results in the prediction of the EAD. A variation of the time period used prior to default other than the cohort method would also be an interesting extension. Also, previous work stated in the literature review section has already looked at some alternative techniques, such as a generalised beta link model. A benchmarking study including this and the techniques mentioned in this chapter may give a better understanding of any improvements that could be made over an ordinary least squares regression model or the logistic regression models suggested in this chapter. The availability of application data in the modelling process may also provide some additional predictive variables in the modelling of the CCF.

Chapter 7

7 Conclusions

In this PhD thesis, we addressed three issues relating to the implementation of the advanced internal ratings based approach (AIRB) by financial institutions. The issues raised in this thesis included that of building classification models for the estimation of probability of default (PD) for imbalanced credit scoring data sets; the accurate prediction of loss given default (LGD); and the construction of a robust and comprehensible model for exposure at default (EAD).

In this chapter we display the conclusions that can be drawn from the research undertaken in this thesis. After highlighting the conclusions from each project, issues for further research will also be given.

7.1 Thesis Summary and Conclusions

In the literature review of this thesis (cf. Chapter 2), we identified issues pertaining to the estimation of probability of default (PD) in imbalanced credit scoring data sets. Although to date a lot of work has been undertaken in the field of PD estimation, the issue of imbalanced data sets has as of yet not been fully addressed.
In Chapter 4 of this thesis, we addressed this issue of estimating probability of default for imbalanced data sets. We achieved this by looking at a number of credit scoring techniques, and studying their performance over various class distributions on five real-life credit data sets. Two techniques that have yet to be fully researched in the context of credit scoring, i.e. Gradient Boosting and Random Forests, were also chosen to give a broader review of the techniques available. The classification power of these techniques was assessed based on the area under the receiver operating

characteristic curve (AUC). Friedman's test and Nemenyi's post-hoc tests were then applied to determine whether the differences between the average ranked performances of the AUCs were statistically significant. Finally, these significance results were visualised using significance diagrams for each of the various class distributions analysed.

The results of these experiments showed that the Gradient Boosting and Random Forest classifiers performed well in dealing with samples where a large class imbalance was present. It does appear that in extreme cases the ability of random forests and gradient boosting to concentrate on 'local' features in the imbalanced data is useful. The most commonly used credit scoring techniques, linear discriminant analysis (LDA) and logistic regression (LOG), gave results that were reasonably competitive with the more complex techniques and this competitive performance continued even when the samples became much more imbalanced. This would suggest that the currently most popular approaches are fairly robust to imbalanced class sizes. On the other hand, techniques such as QDA and C4.5 were significantly worse than the best performing classifiers. It can also be concluded that the use of a linear kernel LS-SVM would not be beneficial in the scoring of data sets where a very large class imbalance exists.

The second major issue identified in the implementation of an advanced internal ratings based approach is the estimation of loss given default (LGD). To address this issue, in Chapter 5, a large scale LGD study evaluating 17 regression techniques on 6 real life lending data sets from major international banking institutions was undertaken. The average predictive performance of the models in terms of R^2 ranges from 4 % to 43 %, which indicates that most resulting models have limited explanatory power. Nonetheless, a clear trend can be seen that non-linear techniques and artificial neural networks and support vector machines in particular give higher performances than more traditional linear techniques. This indicates the presence of non-linear interactions between the independent variables and the LGD, contrary to some studies in PD modelling where the difference between linear and non-linear techniques is not that explicit. Given the fact that LGD has a bigger impact on the minimal capital requirements than PD, we demonstrated the potential and importance

of applying non-linear techniques, preferably in a two-stage context to obtain comprehensibility as well, for LGD modelling.

Finally the issue of regression model development for credit card exposure at default (EAD) is dealt with in Chapter 6 of this thesis. This chapter sets out with the aim of developing a comprehensible and robust regression model for the estimation of Exposure at Default (EAD) for consumer credit cards through the prediction of the credit conversion factor (CCF). An in-depth analysis of the predictive variables used in the modelling of the CCF is also given, showing that previously acknowledged variables are significant and identifying a series of additional variables.
The results from this chapter show that a marginal improvement in the coefficient of determination can be achieved with the use of a binary logit model over a traditional OLS model. Interestingly the use of a cumulative logit model performs worse than both the binary logit and OLS models. The probable cause of this are the size of the peaks around 0 and 1 compared to the number of observations found in the interval between the two peaks. This therefore allows for more error in the prediction of the CCF via a cumulative three-class model.
Another interesting finding is that although the predictive power of the CCF is weak, when this predicted value is applied to the EAD formulation to predict the actual EAD value, the predictive power is fairly strong. In particular when the predictive values obtained through the application of the OLS with Beta transformation model were applied to the EAD formulation an improvement in the coefficient of determination was seen. Nonetheless, similar performance, in terms of correlations, could be achieved by a simple model that takes the average CCF of the previous cohort, showing that much of the explanatory power of EAD modelling derives from the current exposure.
With regards to the additional variables proposed in the prediction of the CCF only one, i.e. average number of days delinquent in the last 6 months, gave an adequate p-value, whilst undrawn percentage, potentially an alternative to credit percentage, was significant for the OLS with Beta transformation model. Even though the relative changes in the undrawn amount give reasonable information value scores, these variables do not prove to be significant in the regression models, probably due to their high correlation with the undrawn variable. This shows that the actual values at the

start of the cohort already give a significant representation of previous activity in order to predict the CCF.

In summary, this thesis has identified and presented detailed results and findings for three main issues facing financial institutions wishing to implement an AIRB approach. An extensive review of the current literature and findings has also been presented and extrapolated upon with the aim of presenting a better understanding for financial institutions considering appropriate techniques and methodologies in the modelling process.

7.2 Issues for further research

Further to the conclusions presented in this thesis, there still remain many challenging issues for further research. This section will highlight the issues for further research identified by each of the major works conducted in this thesis.

7.2.1 The imbalanced data set problem

With regards to probability of default (PD) modelling for imbalanced data sets further work that could be conducted, as a result of the findings presented in this thesis, would be to firstly consider a stacking approach to classification through the combination of multiple techniques. Such an approach would allow a meta-learner to pick the best model to classify an observation. Secondly, another interesting extension to the research would be to apply these techniques on much larger data sets which display a wider variety of class distributions. It would also be of interest to look into the effect of not only the percentage class distribution but also the effect of the actual number of observations in a data set.
Finally, as stated in the literature review chapter of this thesis, there have been several approaches already researched in the area of oversampling techniques to deal with large class imbalances, in the area of machine learning. Further research into this and their effect on credit scoring model performance would be beneficial.

7.2.2 Loss Given Default

In the literature to date there has been considerable evidence that macroeconomic factors affect a client's credit risk behaviour. To further the research presented in this thesis it maybe a worthwhile endeavour to investigate the influence of macro-economic variables, both in the context of improving LGD models and for stress testing.

A variety of LGD data sets have been analysed and reported in Chapter 5 of this thesis. To further this work separate studies on corporate and retail credit LGD data sets could be made, to determine whether separate risk drivers are present in the prediction of each. Finally, one could also try to add comprehensibility to well-performing black box models with rule extraction techniques to gain more insight.

7.2.3 Exposure at Default

There is an obvious need for further research into the prediction of the exposure at default (EAD) value as this thesis can only go so far in its estimations. A more extensive study with multiple data sets over a longer timescale would be able to give more reliable results in the prediction of the EAD. A variation of the time period used prior to default other than the cohort method would also be an interesting extension. Also, previous work stated in the literature review section has already looked at some alternative techniques, such as a generalised beta link model. A benchmarking study including this and the techniques mentioned in this thesis may give a better understanding of any improvements that could be made over an ordinary least squares regression model or the logistic regression models suggested in this thesis. The availability of application data in the modelling process may also provide some additional predictive variables in the modelling of the CCF.

Appendices

A1: Data sets used in Chapter 4

A1.1 Australian Credit

THIS CREDIT DATA ORIGINATES FROM QUINLAN (see below).

1. Title: Australian Credit Approval
2. Sources: quinlan@cs.su.oz.au
3. This file concerns credit card applications. All attribute names and values have been changed to meaningless symbols to protect confidentiality of the data.
4. Number of Attributes: 14 + class attribute
5. Class Distribution:

   +: 307 (44.5%) CLASS 2

   -: 383 (55.5%) CLASS 1

Variable name	Type
A1	Nominal
A2	Continuous
A3	Continuous
A4	Nominal
A5	Nominal
A6	Nominal
A7	Continuous

A8	Nominal
A9	Nominal
A10	Continuous
A11	Nominal
A12	Nominal
A13	Continuous
A14	Continuous
A15	Binary Target

A1.2 Bene1

Variable name	Type
Identification number	continuous
Amount of loan	continuous
Amount on purchase invoice	continuous
Percentage of financial burden	continuous
Term	continuous
Personal loan	nominal
Purpose	nominal
Private or professional loan	nominal
Monthly payment	continuous
Savings account	continuous
Other loan expenses	continuous
Income	continuous
Profession	nominal

Number of years employed	continuous
Number of years in Belgium	continuous
Age	continuous
Applicant Type	nominal
Nationality	nominal
Marital status	nominal
Number of years since last house move	continuous
Code of regular saver	nominal
Property	nominal
Existing credit info	nominal
Number of years client	continuous
Number of years since last loan	continuous
Number of checking accounts	continuous

A1.3 Bene2

The variable names for the Bene2 dataset cannot be displayed for confidentiality purposes. The dataset includes:

28 input variables:

- Continuos variables: 18
- Nominal variables: 10

1 Binary class variable

A1.4 Behav

The variable names for the Behav dataset cannot be displayed for confidentiality purposes. The dataset includes:

1 ID variable

60 Input variables:

- Nominal variables: 10
- Ordinal variables: 1
- Continuous variables: 49

1 Binary class variable (0 = “good account”; 1 = “bad account”)

A1.5 German Credit

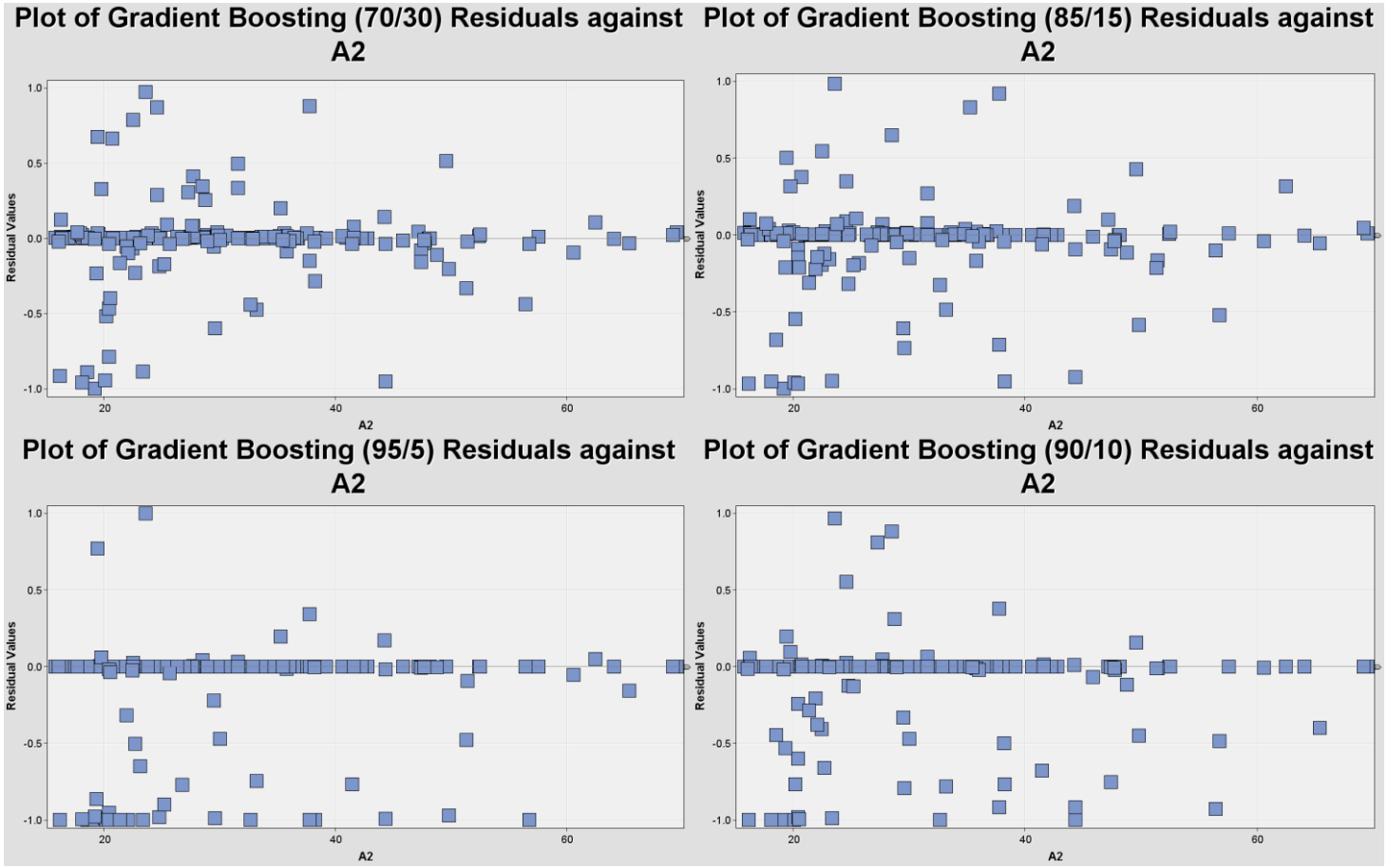
Vaiable name	**Type**
Checking Status	Nominal
Duration	Continuous
Credit History	Nominal
Purpose	Nominal
Credit_Amount	Continuous
Savings Status	Nominal
Employment	Nominal
Installment_commitment	Continuous
Personal Status	Nominal
Other_Parties	Nominal
Residence_Since	Continuous
Property	Nominal

Age	Continuous
Other_Payment_Plans	Nominal
Housing	Nominal
Existing Credits	Continuous
Job	Nominal
Number of Dependents	Continuous
Own_Telephone	Nominal
Foreign_worker	Nominal

A2: Residual plots for Chapter 4

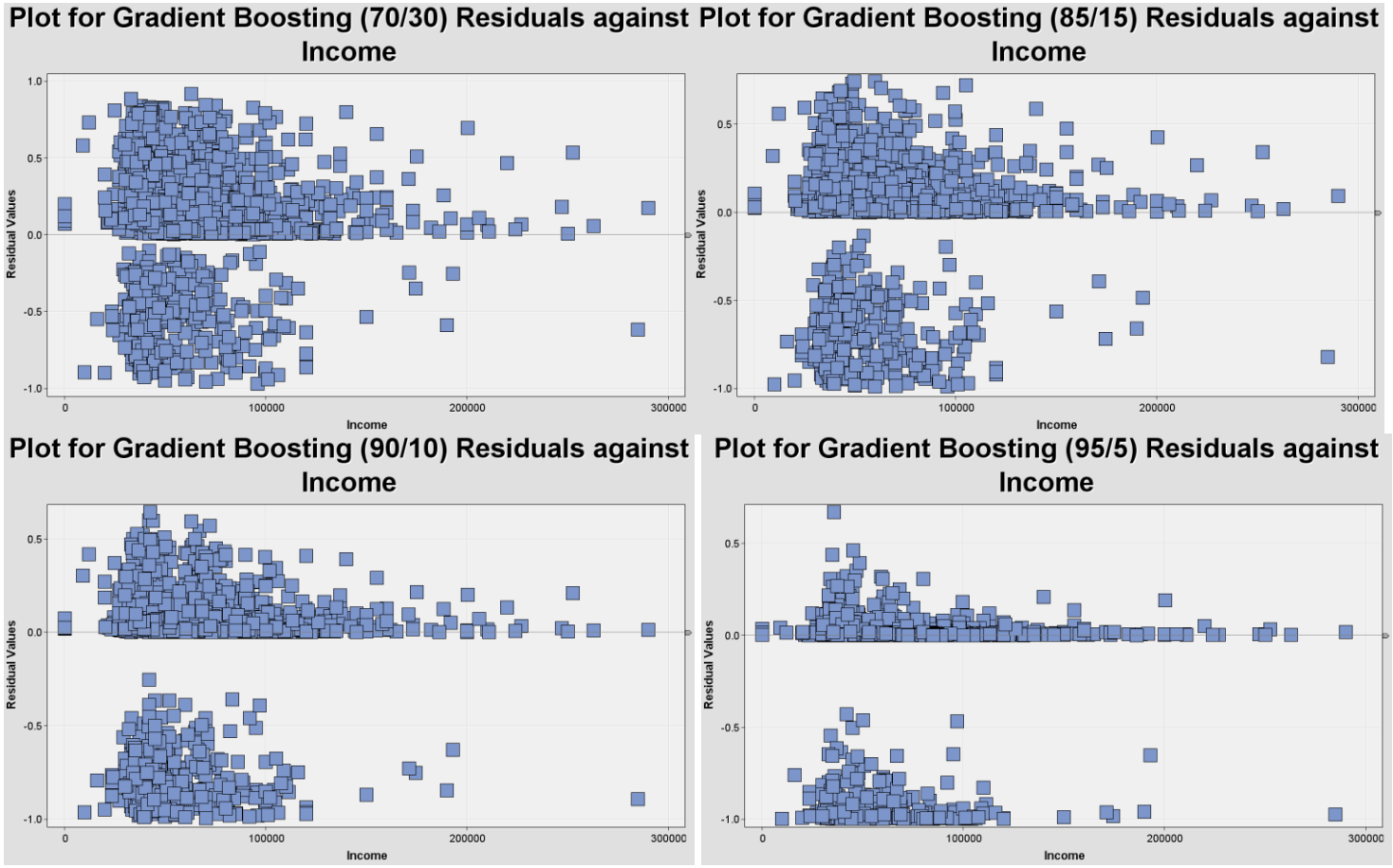
A2.1 Australian Credit: Gradient Boosting

The following plots show the residual values of the Gradient Boosting classifier over varying class imbalances of the Australian Credit dataset against the A2 variable. It can be seen that as the class imbalance increases the larger the concentration of negative residuals are present.


Plot of Gradient Boosting (70/30) Residuals against A2
Plot of Gradient Boosting (85/15) Residuals against A2
Plot of Gradient Boosting (95/5) Residuals against A2
Plot of Gradient Boosting (90/10) Residuals against A2
Residual Values
A2
1.0
0.5
0.0
-0.5
-1.0
20
40
60

A2.2 Bene2: Gradient Boosting

The following plots show the residual values of the Gradient Boosting classifier over varying class imbalances of the Bene2 dataset against the Income variable. It can be seen that as the class imbalance increases the larger the concentration of negative one residuals are present.



A3: Stepwise variable selection for Linear models used in Chapter 5

A3.1 BANK1

Summary of Stepwise Selection							
Step	Variable Entered	Number Vars In	Partial R-Square	Model R-Square	C(p)	F Value	Pr > F
1	AGE_OF_EXP	1	0.0191	0.0191	2702.63	620.14	<.0001
2	MTHS_ARRS_ADV	2	0.0133	0.0324	2235.62	438.33	<.0001
3	no_mths_arrs_0_12m	3	0.0182	0.0506	1594.58	612.50	<.0001
4	LOAN_AMT_LAST	4	0.0078	0.0583	1322.97	262.76	<.0001
5	APP_SCORE_FIRST	5	0.0072	0.0656	1070.33	246.42	<.0001
6	Worst_arrs_12m	6	0.0073	0.0729	813.798	252.15	<.0001
7	JOINT_APP	7	0.0032	0.0761	701.308	112.05	<.0001
8	no_mths_arrs_0_ever	8	0.0029	0.0790	602.210	99.25	<.0001
9	TERM_LAST	9	0.0035	0.0825	479.866	122.54	<.0001
10	TADD	10	0.0020	0.0845	412.281	68.72	<.0001
11	TIME_AT_BANK	11	0.0011	0.0856	376.712	37.14	<.0001
12	RESID_STATUS_FIRST2	12	0.0011	0.0867	339.973	38.35	<.0001
13	EMPL_STATUS_C1_FIRST4	13	0.0012	0.0879	299.319	42.28	<.0001
14	Worst_arrs_6m	14	0.0007	0.0886	275.080	26.03	<.0001
15	HBS_MORT_HELD_FIRST	15	0.0007	0.0893	253.908	23.00	<.0001
16	EMPL_STATUS_C1_FIRST8	16	0.0006	0.0899	234.427	21.34	<.0001
17	PL_HELD_FIRST	17	0.0006	0.0905	216.190	20.11	<.0001

Summary of Stepwise Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
18	no_mths_arrs_2_ever	18	0.0005	0.0909	201.539	16.56	<.0001
20	no_mths_arrs_2_12m	18	0.0011	0.0918	170.520	37.40	<.0001
21	Worst_arrs_ever	19	0.0005	0.0923	154.726	17.72	<.0001
22	Max_con_mths_arrs_1	20	0.0005	0.0928	139.693	16.97	<.0001
23	EMPL_STATUS_C1_FIRST3	21	0.0004	0.0932	126.855	14.79	0.0001
24	no_mths_arrs_0_3m	22	0.0004	0.0936	115.057	13.76	0.0002
25	no_mths_arrs_2_6m	23	0.0005	0.0941	100.934	16.08	<.0001
26	PURP2	24	0.0003	0.0944	92.4117	10.50	0.0012
27	MAXIM_HELD_FIRST	25	0.0004	0.0947	81.7724	12.62	0.0004
28	EMPL_STATUS_C1_FIRST10	26	0.0003	0.0950	73.5318	10.23	0.0014
29	EMPL_STATUS_C1_FIRST6	27	0.0003	0.0953	65.6508	9.87	0.0017
30	Worst_arrs_3m	28	0.0002	0.0955	60.7246	6.92	0.0085

A3.2 BANK2

Summary of Stepwise Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	arr_perc_UOS	1	0.1575	0.1575	8616.94	14859.5	<.0001
2	security5	2	0.0295	0.1870	5538.31	2880.03	<.0001
3	ltv	3	0.0182	0.2052	3639.08	1818.08	<.0001
4	TOB_UOS	4	0.0140	0.2191	2180.01	1422.15	<.0001
5	propage2	5	0.0069	0.2261	1457.78	711.24	<.0001
6	region12	6	0.0027	0.2288	1177.53	278.15	<.0001
7	arrmths_def	7	0.0018	0.2306	987.637	189.56	<.0001
8	security3	8	0.0016	0.2322	819.708	168.21	<.0001
9	region10	9	0.0018	0.2340	630.777	189.45	<.0001
10	bal_def	10	0.0012	0.2353	505.006	126.98	<.0001
11	orig_loan	11	0.0016	0.2369	340.993	165.33	<.0001
12	term	12	0.0009	0.2378	246.414	96.30	<.0001
13	propage4	13	0.0006	0.2384	184.566	63.71	<.0001
14	endowment	14	0.0004	0.2388	145.235	41.26	<.0001
15	region5	15	0.0003	0.2391	119.039	28.16	<.0001
16	region8	16	0.0003	0.2393	92.8840	28.13	<.0001
17	security1	17	0.0002	0.2396	70.9766	23.89	<.0001
18	region2	18	0.0002	0.2398	50.4508	22.52	<.0001
19	region6	19	0.0001	0.2399	37.2918	15.16	<.0001
20	region11	20	0.0001	0.2400	25.7214	13.57	0.0002
21	region4	21	0.0001	0.2401	20.8574	6.86	0.0088

A3.3 BANK3

Summary of Stepwise Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	QUOTITEIT1	1	0.0110	0.0110	40.9197	24.78	<.0001
2	EAD	2	0.0082	0.0192	24.1071	18.64	<.0001
3	LTV_RAT	3	0.0063	0.0255	11.6560	14.40	0.0002
4	PROVINCIE_PAND4	4	0.0041	0.0296	4.1513	9.51	0.0021

A3.4 BANK4

Summary of Stepwise Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	Utilization	1	0.3589	0.3589	990.479	2943.60	<.0001
2	PD_Rnd	2	0.0752	0.4341	259.828	698.53	<.0001
3	LTV	3	0.0198	0.4539	69.2096	190.26	<.0001
4	Age	4	0.0029	0.4567	43.1425	27.86	<.0001
5	PrinBal	5	0.0039	0.4606	7.2950	37.84	<.0001

A3.5 BANK5

Summary of Stepwise Selection							
Step	**Variable Entered**	**Number Vars In**	**Partial R-Square**	**Model R-Square**	**C(p)**	**F Value**	**Pr > F**
1	amount_funded	1	0.1068	0.1068	691.938	326.41	<.0001
2	Total_Debt	2	0.0283	0.1350	585.683	89.21	<.0001
3	occupancy_code_description1	3	0.0242	0.1593	494.881	78.65	<.0001
4	product_family_name1	4	0.0283	0.1876	388.438	95.08	<.0001
5	amortization_type3	5	0.0176	0.2051	323.198	60.23	<.0001
6	Original_Appraised_Value	6	0.0154	0.2205	266.220	53.86	<.0001
7	jumbo_indicator	7	0.0119	0.2324	222.573	42.32	<.0001
8	State4	8	0.0105	0.2430	184.196	37.94	<.0001
9	business_line_crm_new3	9	0.0073	0.2503	158.381	26.38	<.0001
10	loan_purpose_type2	10	0.0069	0.2572	133.945	25.29	<.0001
11	Loan_Category2	11	0.0043	0.2615	119.370	15.95	<.0001
12	State3	12	0.0040	0.2655	106.120	14.75	0.0001
13	total_debt_ratio	13	0.0041	0.2696	92.3181	15.36	<.0001
14	property_type_description2	14	0.0022	0.2718	85.9808	8.13	0.0044

A3.6 BANK6

The variable names for the BANK6 dataset cannot be displayed for confidentiality purposes.

A4: R-square based variable selection for Non-linear used in Chapter 5

The R-Square selection process implements a forward stepwise least squares regression that maximizes the model R-square value.

A4.1 BANK1

Effects Chosen for Target: LGD

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: AGE_OF_EXP	1	0.019068	620.142935	<.0001	93.422609	0.150647
Var: LOAN_AMT_LAST	1	0.010770	354.154211	<.0001	52.768110	0.148998
Var: APP_SCORE_FIRST	1	0.008439	279.931233	<.0001	41.347551	0.147706
Var: no_mths_arrs_0_12m	1	0.004723	157.417922	<.0001	23.138136	0.146985
Var: no_mths_arrs_0_ever	1	0.011344	382.653388	<.0001	55.579504	0.145248
Class: JOINT_APP	1	0.002996	101.364291	<.0001	14.676747	0.144792

A4.2 BANK2

Effects Chosen for Target: LGD_UOS

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: arr_perc_UOS	1	0.157521	14860	<.0001	422.449567	0.028428
Class: security	4	0.036218	892.506531	<.0001	97.132223	0.027208
Var: TOB_UOS	1	0.017561	1769.509728	<.0001	47.096270	0.026615
Var: ltv	1	0.012248	1253.560225	<.0001	32.846357	0.026202
Class: region	12	0.005642	48.470562	<.0001	15.132107	0.026016
Class: propage	4	0.005388	139.815495	<.0001	14.448777	0.025835
Var: arrmths_def	1	0.001664	173.075664	<.0001	4.461822	0.025780
Var: bal_def	1	0.001040	108.339709	<.0001	2.789187	0.025745
Var: orig_loan	1	0.001631	170.264701	<.0001	4.374116	0.025690
Var: term	1	0.000873	91.288349	<.0001	2.342544	0.025661

A4.3 BANK3

Effects Chosen for Target: lgd

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: EAD	1	0.013908	31.453063	<.0001	0.694248	0.022073
Var: SALDO_VOORG	1	0.011851	27.113396	<.0001	0.591534	0.021817
Var: QUOTITEIT1	1	0.009552	22.060941	<.0001	0.476799	0.021613

A4.4 BANK4

Effects Chosen for Target: LGD

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: Utilization	1	0.358906	2943.603542	<.0001	408.015042	0.138611
Var: PD_Rnd	1	0.075194	698.525201	<.0001	85.482940	0.122376
Var: LTV	1	0.019769	190.257422	<.0001	22.473918	0.118124
Var: Age	1	0.002881	27.864848	<.0001	3.274762	0.117523
Var: PrinBal	1	0.003884	37.838227	<.0001	4.415908	0.116705

A4.5 BANK5

Effects Chosen for Target: LGD

Effect	DF	R-Square	F Value	p-Value	Sum of Squares	Error Mean Square
Var: amount_funded	1	0.106760	326.409796	<.0001	21.201256	0.064953
Class: product_family_name	3	0.032109	33.906469	<.0001	6.376473	0.062687
Class: occupancy_code_description	2	0.029295	48.000950	<.0001	5.817597	0.060599
Var: Total_Debt	1	0.025256	85.326501	<.0001	5.015530	0.058780
Class: State	3	0.016939	19.463418	<.0001	3.363830	0.057609
Class: amortization_type	2	0.016300	28.664730	<.0001	3.236922	0.056462
Var: amount_appraised	1	0.013089	46.813466	<.0001	2.599388	0.055527
Class: jumbo_indicator	1	0.009907	35.886234	<.0001	1.967395	0.054823
Class: business_line_crm_new	5	0.010578	7.758844	<.0001	2.100700	0.054150
Class: loan_purpose_type	3	0.004816	5.919241	0.0005	0.956375	0.053857
Var: total_debt_ratio	1	0.004148	15.376455	<.0001	0.823758	0.053573
Class: property_type_description	6	0.003404	2.108120	0.0493	0.675967	0.053442
Class: Loan_Type	9	0.002347	0.968925	0.4637	0.466076	0.053447
Class: ARM_Indicator	1	0.002246	8.369912	0.0038	0.446126	0.053301
Class: primary_borrower_self_employed	2	0.001936	3.614265	0.0271	0.384543	0.053198
Var: Annual_Interest_Rate	1	0.002000	7.484548	0.0063	0.397205	0.053070
Class: firsttime_homebuyer_indicator	2	0.001864	3.493172	0.0305	0.370080	0.052972
Var: Unpaid_Interest	1	0.001649	6.193721	0.0129	0.327460	0.052870
Var: Original_Appraised_Value	1	0.001639	6.166712	0.0131	0.325406	0.052768
Var: score_fico_used	1	0.000725	2.731085	0.0985	0.144022	0.052734

A4.6 BANK6

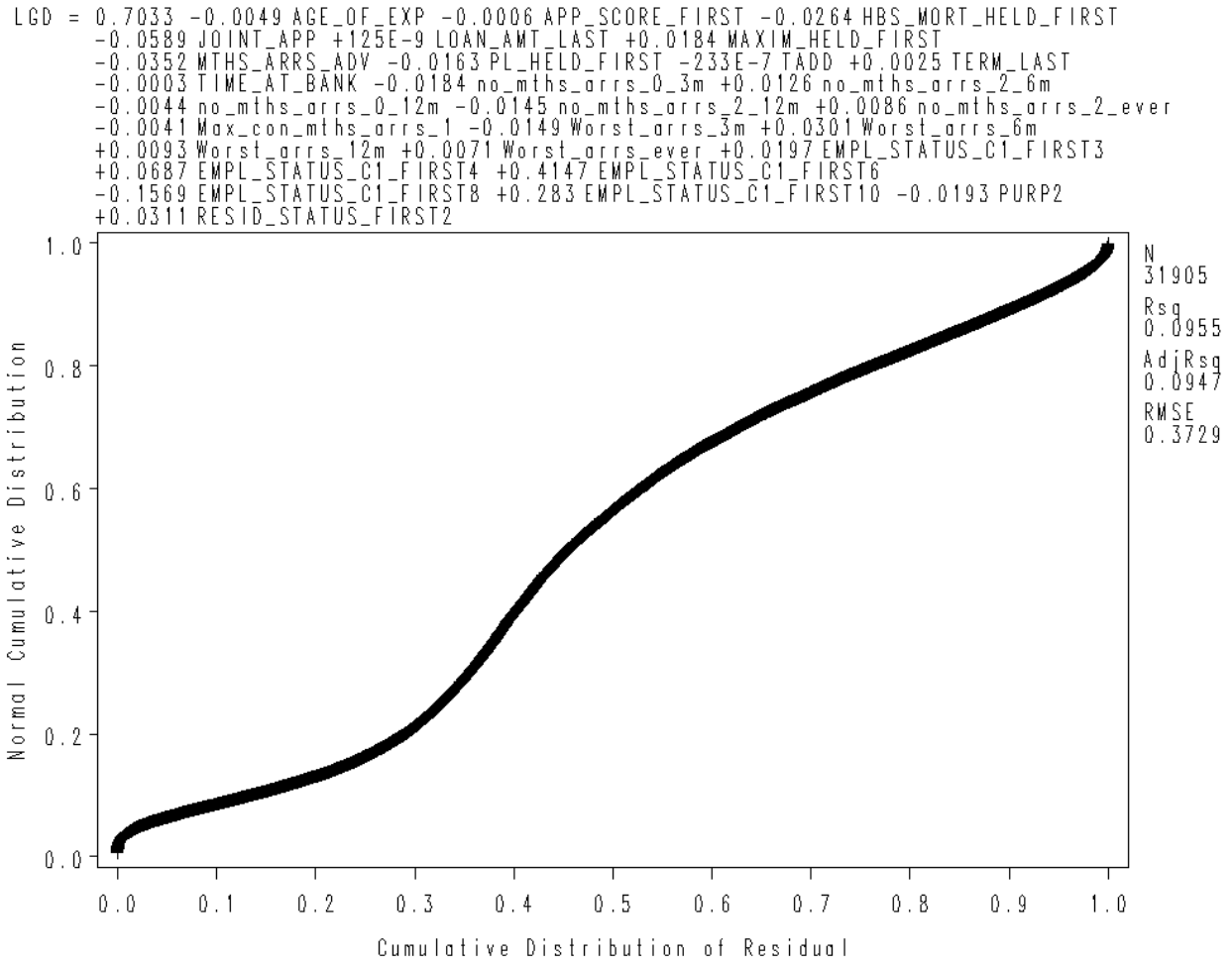
The variable names for the BANK6 dataset cannot be displayed for confidentiality purposes.

A5: Normal probability plots for techniques used in Chapter 5

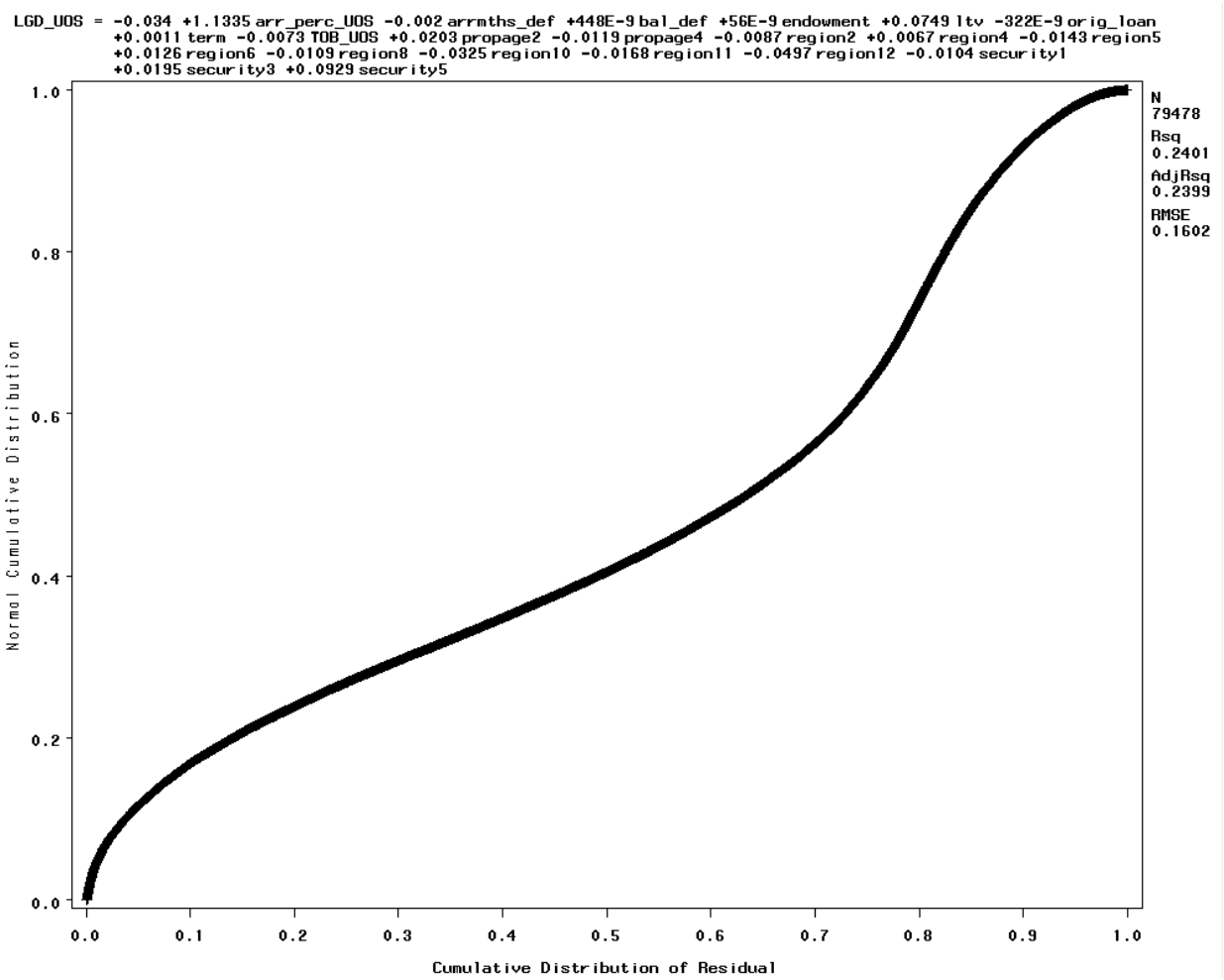
The following plots detail the normal cumulative distribution vs. the cumulative distribution of residuals for the OLS regression model over the six data sets analysed in this thesis. All the data sets (BANK1, BANK2, BANK3, BANK4 and BANK6) apart from potentially BANK5 do not

display diagonal normal probability. It therefore seems that the normality assumption is not satisfied for these data sets, leading to the summation that the OLS model fit is relatively poor.

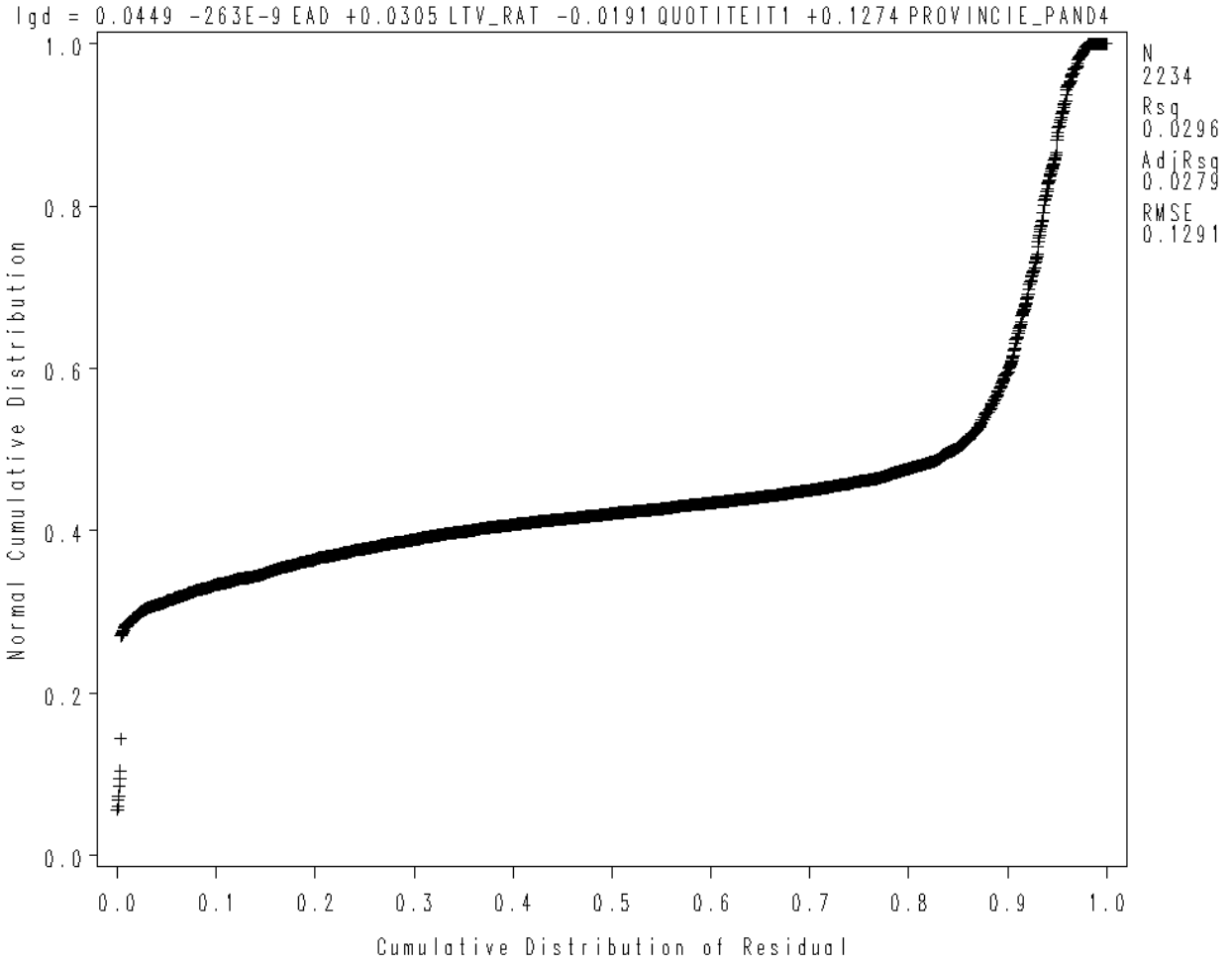
A5.1 BANK1 OLS model normal probability plots



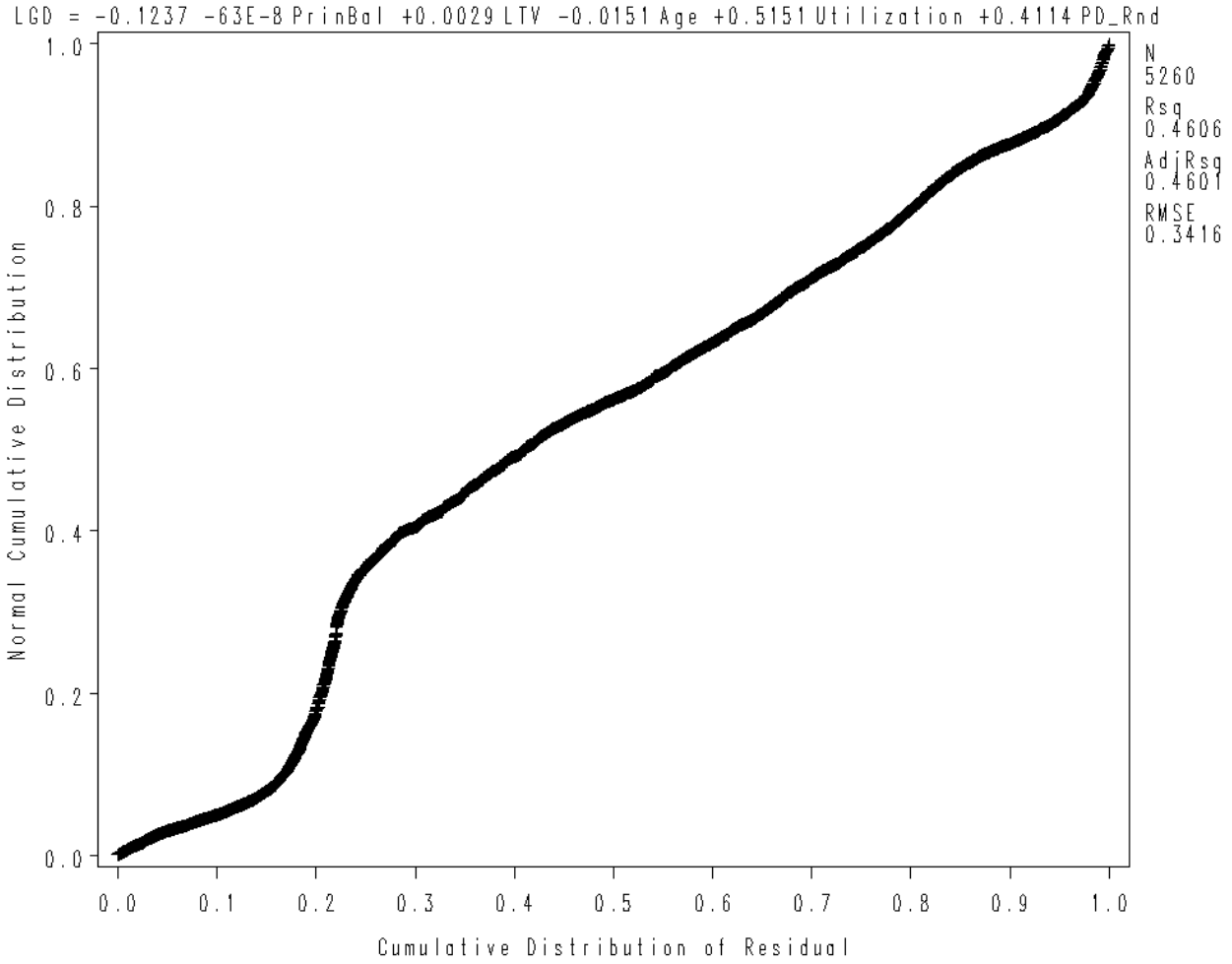
A5.2 BANK2 OLS model normal probability plots



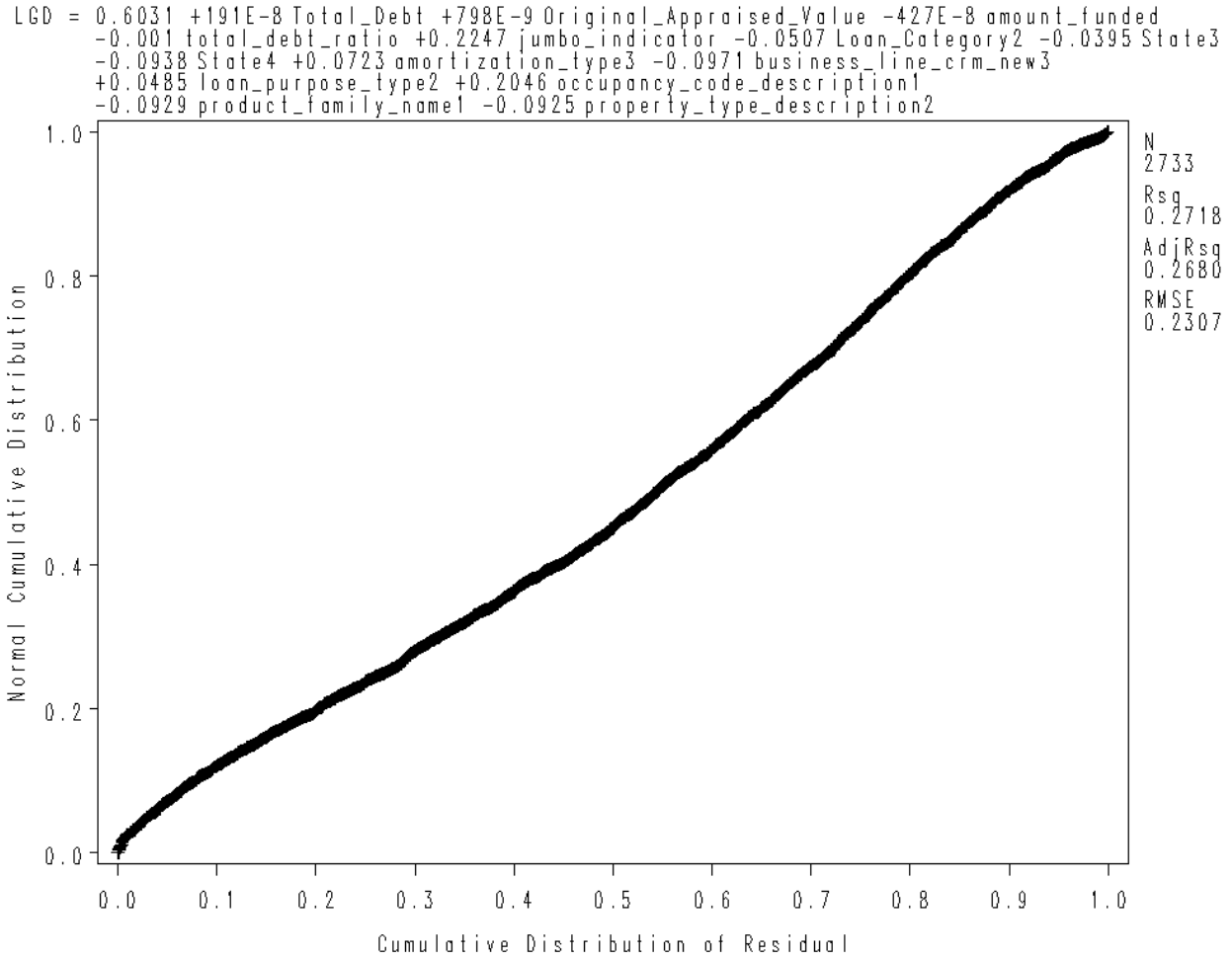
A5.3 BANK3 OLS model normal probability plots



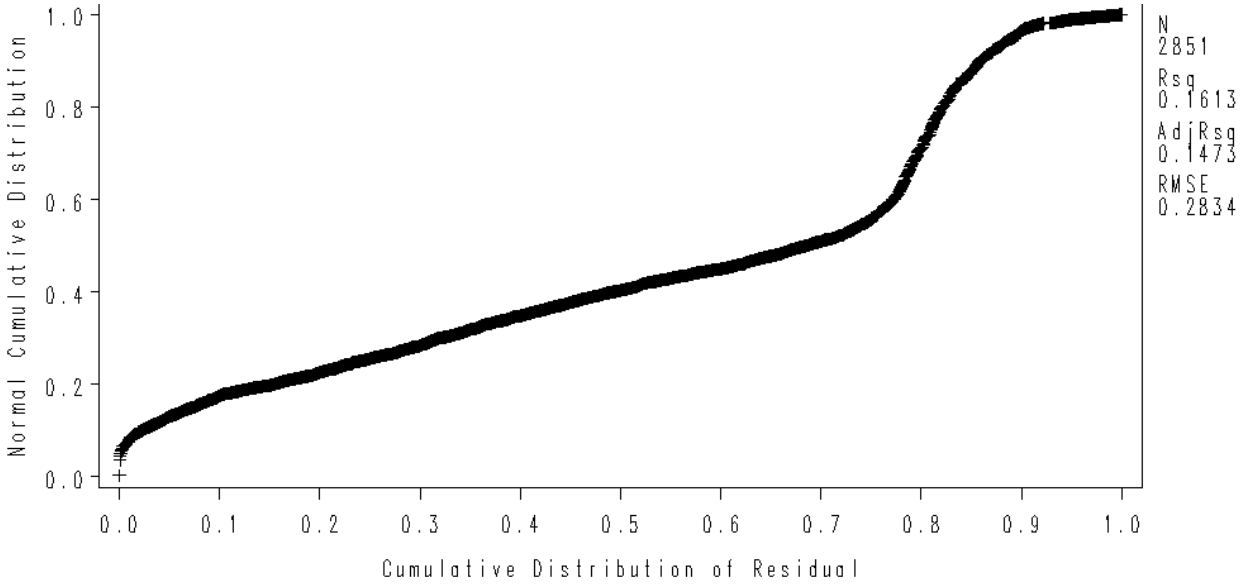
A5.4 BANK4 OLS model normal probability plots



A5.5 BANK5 OLS model normal probability plots



A5.6 BANK6 OLS model normal probability plots



A6: Pearson's correlation coefficients matrix for input variables used in Chapter 6

	CCF	EAD	Commit_ Amt	Drawn _Amt	Undrawn_ Amt	Credit %	Time_d efault	Rating _grade _1	Rating _grade _2	Rating _grade _3	Rating _grade _4	Av_No_d ays_del_3	Av_No_d ays_del_6	Av_No_d ays_del_9	Av_No_d ays_del_1 2	Incr_com mit_Amt
EAD	0.323	1.000														
Commit_Amt	0.030	0.712	1.000													
Drawn_Amt	0.089	0.755	0.782	1.000												
Undrawn_Amt	-0.083	0.012	0.421	-0.236	1.000											
Credit%	0.039	0.212	-0.067	0.458	-0.771	1.000										
Time_default	0.229	0.078	0.046	-0.021	0.102	-0.111	1.000									
Rating_grade_1	-0.041	-0.050	0.077	-0.136	0.318	-0.319	-0.004	1.000								
Rating_grade_2	0.231	0.121	0.177	0.095	0.138	-0.114	0.280	-0.214	1.000							
Rating_grade_3	-0.093	-0.038	-0.106	0.007	-0.176	0.185	-0.106	-0.118	-0.654	1.000						
Rating_grade_4	-0.184	-0.094	-0.154	-0.068	-0.141	0.098	-0.253	-0.084	-0.466	-0.258	1.000					
Av_No_days_del_3	-0.067	-0.072	-0.099	-0.038	-0.099	0.112	-0.134	-0.097	-0.404	0.132	0.447	1.000				
Av_No_days_del_6	-0.053	-0.072	-0.096	-0.039	-0.093	0.103	-0.103	-0.107	-0.397	0.160	0.407	0.844	1.000			
Av_No_days_del_9	-0.056	-0.084	-0.105	-0.054	-0.085	0.087	-0.089	-0.111	-0.391	0.175	0.382	0.757	0.925	1.000		
Av_No_days_del_12	-0.050	-0.085	-0.109	-0.060	-0.083	0.080	-0.076	-0.113	-0.386	0.178	0.373	0.719	0.874	0.961	1.000	
Incr_commit_Amt	0.070	0.293	0.349	0.286	0.128	-0.021	0.080	-0.020	0.223	-0.089	-0.188	-0.127	-0.160	-0.195	-0.211	1.000
Undrawn%	-0.039	-0.212	0.067	-0.458	0.771	-1.000	0.111	0.319	0.114	-0.185	-0.098	-0.112	-0.103	-0.087	-0.080	0.021
Abs_change_drawn_3	0.036	0.260	0.222	0.414	-0.256	0.263	-0.072	-0.055	0.013	0.020	-0.014	-0.090	-0.086	-0.084	-0.077	0.142
Abs_change_drawn_6	0.019	0.326	0.277	0.482	-0.270	0.308	-0.078	-0.071	0.023	0.015	-0.013	-0.060	-0.099	-0.106	-0.105	0.239
Abs_change_drawn_12	0.053	0.439	0.374	0.595	-0.281	0.359	-0.068	-0.105	0.031	0.022	-0.014	-0.051	-0.074	-0.094	-0.105	0.297
Abs_change_undrawn_3	-0.008	-0.167	-0.082	-0.328	0.349	-0.308	0.117	0.056	0.075	-0.076	-0.040	0.033	0.030	0.027	0.020	-0.001
Abs_change_undrawn_6	0.008	-0.183	-0.078	-0.357	0.398	-0.368	0.119	0.075	0.097	-0.087	-0.066	-0.022	0.012	0.017	0.018	0.029
Abs_change_undrawn_12	-0.017	-0.236	-0.091	-0.411	0.456	-0.430	0.124	0.113	0.110	-0.100	-0.088	-0.028	-0.016	-0.009	-0.002	0.128
Abs_change_commit_3	0.061	0.202	0.302	0.185	0.202	-0.095	0.097	0.001	0.190	-0.120	-0.115	-0.125	-0.122	-0.125	-0.123	0.305
Abs_change_commit_6	0.051	0.291	0.390	0.269	0.216	-0.086	0.070	0.002	0.223	-0.131	-0.148	-0.156	-0.168	-0.173	-0.170	0.515
Abs_change_commit_12	0.061	0.364	0.486	0.345	0.255	-0.083	0.084	0.003	0.228	-0.125	-0.163	-0.130	-0.151	-0.174	-0.182	0.706
Rel_change_drawn_3	0.013	-0.038	-0.035	-0.049	0.016	-0.022	0.015	0.006	-0.018	0.009	0.011	0.007	0.008	0.009	-0.005	-0.020
Rel_change_drawn_6	-0.024	-0.026	-0.035	-0.048	0.015	-0.056	0.015	0.014	-0.016	0.016	-0.004	-0.012	-0.009	-0.002	0.002	-0.028
Rel_change_drawn_12	-0.002	0.007	0.010	0.014	-0.004	-0.010	-0.027	0.004	-0.038	-0.007	0.059	0.015	0.006	0.009	0.005	0.001
Rel_change_undrawn_3	-0.013	-0.002	-0.005	0.003	-0.012	0.025	-0.001	0.002	0.019	-0.038	0.019	0.013	0.008	0.008	0.009	-0.016
Rel_change_undrawn_6	-0.004	-0.008	0.005	-0.013	0.026	-0.013	-0.009	0.000	0.008	-0.007	-0.002	0.002	-0.001	0.007	0.006	0.000
Rel_change_undrawn_12	-0.006	-0.015	-0.014	-0.030	0.022	-0.026	-0.010	0.028	0.012	-0.025	-0.001	0.008	-0.002	-0.002	0.000	0.012
Rel_change_commit_3	0.074	0.007	0.021	-0.028	0.073	-0.104	0.096	-0.020	0.144	-0.064	-0.109	-0.119	-0.117	-0.115	-0.115	0.257
Rel_change_commit_6	0.061	0.017	0.007	-0.030	0.055	-0.088	0.080	-0.018	0.166	-0.079	-0.122	-0.138	-0.157	-0.155	-0.153	0.448
Rel_change_commit_12	0.062	0.029	0.013	-0.018	0.047	-0.076	0.075	-0.017	0.172	-0.077	-0.134	-0.112	-0.140	-0.156	-0.163	0.600

	Undrawn %	Abs_cha nge_dra wn_3	Abs_cha nge_dra wn_6	Abs_cha nge_dra wn_12	Abs_chan ge_undra wn_3	Abs_chan ge_undra wn_6	Abs_chan ge_undra wn_12	Abs_chan ge_commi t_3	Abs_chan ge_commi t_6	Abs_chan ge_commi t_12	Rel_chan ge_drawn _3	Rel_chan ge_drawn _6	Rel_chan ge_drawn _12	Rel_chan ge_undra wn_3
Undrawn%	1.000													

Abs_change_drawn_3	-0.263	1.000												
Abs_change_drawn_6	-0.308	0.667	1.000											
Abs_change_drawn_12	-0.359	0.546	0.718	1.000										
Abs_change_undrawn_3	0.308	-0.893	-0.568	-0.455	1.000									
Abs_change_undrawn_6	0.368	-0.564	-0.861	-0.581	0.630	1.000								
Abs_change_undrawn_12	0.430	-0.440	-0.564	-0.817	0.521	0.682	1.000							
Abs_change_commit_3	0.095	0.234	0.216	0.200	0.229	0.141	0.174	1.000						
Abs_change_commit_6	0.086	0.243	0.336	0.311	0.071	0.191	0.174	0.678	1.000					
Abs_change_commit_12	0.083	0.218	0.309	0.380	0.067	0.112	0.224	0.616	0.803	1.000				
Rel_change_drawn_3	0.022	-0.118	-0.074	-0.085	0.119	0.069	0.083	0.003	-0.016	-0.010	1.000			
Rel_change_drawn_6	0.056	-0.078	-0.068	-0.058	0.063	0.054	0.043	-0.032	-0.032	-0.030	0.065	1.000		
Rel_change_drawn_12	0.010	0.032	0.077	0.066	-0.038	-0.084	-0.080	-0.014	-0.007	-0.018	0.010	0.083	1.000	
Rel_change_undrawn_3	-0.025	0.021	0.012	-0.009	-0.029	-0.020	0.005	-0.018	-0.013	-0.007	0.000	-0.001	-0.001	1.000
Rel_change_undrawn_6	0.013	-0.039	-0.029	0.004	0.037	0.032	-0.004	-0.003	0.004	-0.001	0.000	0.000	-0.002	0.010
Rel_change_undrawn_12	0.026	-0.010	-0.002	-0.014	0.010	0.003	0.019	-0.001	0.002	0.008	0.001	0.003	0.000	-0.003
Rel_change_commit_3	0.104	0.144	0.100	0.075	0.173	0.140	0.148	0.687	0.451	0.363	0.004	-0.009	-0.005	-0.010
Rel_change_commit_6	0.088	0.137	0.171	0.122	0.067	0.186	0.172	0.442	0.674	0.482	0.001	-0.010	-0.006	-0.007
Rel_change_commit_12	0.076	0.125	0.159	0.182	0.068	0.120	0.214	0.417	0.527	0.651	0.003	-0.009	-0.020	-0.001

	Rel_change_undrawn_6	Rel_change_undrawn_12	Rel_change_commit_3	Rel_change_commit_6	Rel_change_commit_12
Abs_change_drawn_12					
Abs_change_undrawn_3					
Abs_change_undrawn_6					
Abs_change_undrawn_12					
Abs_change_commit_3					
Abs_change_commit_6					
Abs_change_commit_12					
Rel_change_drawn_3					
Rel_change_drawn_6					
Rel_change_drawn_12					
Rel_change_undrawn_3					
Rel_change_undrawn_6	1.000				
Rel_change_undrawn_12	-0.004	1.000			
Rel_change_commit_3	0.003	0.000	1.000		
Rel_change_commit_6	0.006	0.007	0.684	1.000	
Rel_change_commit_12	0.005	0.012	0.546	0.714	1.000

References

Allen L, De Long G and Saunders A (2004). Issues in the credit risk modeling of retail markets, *Journal of Banking & Finance*, **28**(4), 727–752.

Altman E (1968). Financial Ratios, Discriminant Analysis and the Prediction of Corporate Bankruptcy. *Journal of Finance*, **23**(4), 589-609.

Altman E (1989), Measuring Corporate Bond Mortality and Performance, *Journal of Finance*, **44**(4), 909-922.

Altman E (1994). Corporate Distress Diagnosis: Comparisons Using Linear Discriminant Analysis and Neural Networks (the Italian Experience). *Journal of Banking & Finance*, **18**(3), 505-529.

Altman E and Sironi A (2004). Default recovery rates in credit risk modelling: A review of the literature and empirical evidence. Economic Note, **33**, 183–208.

Altman E and Suggitt H J (2000). Default rates in the syndicated bank loan market: A mortality analysis, *Journal of Banking & Finance*, **24**(1-2), 229-253.

Araten M and Jacobs M (2001). Loan Equivalents for Revolving Credits and Advised Lines, *The RMA Journal*, May, 34–39.

Araten M and Keisman D (1994). Portfolio Concentration Model, Chase Manhattan Bank/ Portfolio Analytics Group, Presented to the Risk Management Association (December)

Arminger G, Enache D, and Bonne T (1997). Analyzing Credit Risk Data: A Comparison of Logistic Discrimination, Classification Tree Analysis, and Feedforward Networks, *Computational Statistics*, **12**(2)**,** 293–310.

Asarnow E and Marker J (1995). Historical Performance of the U.S. Corporate Loan Market: 1988–1993. *Journal of Commercial Lending*, **10**(2), 13–32.

Baesens B (2003a). Developing intelligent systems for credit scoring using machine learning techniques, PhD Thesis, Faculty of Economics, KU Leuven.

Baesens B, Setiono R, Mues C and Vanthienen J (2003). Using neural network rule extraction and decision tables for credit-risk evaluation, *Management Science*, **49**(3), 312–329.

Baesens B, Van Gestel T, Viaene S, Stepanova M, Suykens J and Vanthienen J (2003). Benchmarking state of the art classification algorithms for credit scoring. *Journal of the Operational Research Society*, **54**(6), 627-635.

Baesens B, Viaene S, Van Gestel T, Suykens J, Dedene G, De Moor B and Vanthienen J (2000). An empirical assessment of kernel type performance for least squares support vector machine classifiers. *International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies*, **1**, 313-316.

Balthazar L (2004). Pd estimates for Basel II, *Risk*, Apr, 84–85.

Basel Committee on Banking Supervision (2001a). The New Basel Capital Accord, Jan. Available at: http://www.bis.org/publ/bcbsca03.pdf.

Basel Committee on Banking Supervision (2004). International Convergence of Capital Measurement and Capital Standards: a Revised Framework, Bank for International Settlements.

Basel Committee on Banking Supervision (2005). Basel committee newsletter no. 6: validation of low-default portfolios in the Basel II framework, *Technical Report*, Bank for International Settlements.

Bastos J (2010). Forecasting bank loans for loss-given-default. *Journal of Banking & Finance*, **34**(10), 2510-2517.

Bastos J (2010). Predicting bank loan recovery rates with neural networks, *CEMAPRE Working Papers 1003*, Centre for Applied Mathematics and Economics (CEMAPRE), School of Economics and Management (ISEG), Technical University of Lisbon.

Batista G (2004). A Study of the Behavior of Several Methods for Balancing Machine Learning Training Data. *ACM SIGKDD Explorations Newsletter*, **6**(1), 20-29.

Bedingfield S and Smith KA (2001). Evolutionary rule generation and its application to credit scoring. In L. Reznik and V. Kreinovich, editors, *Soft Computing in Measurement and Information Acquisition*, Heidelberg, 2001. Physica-Verlag.

Bellotti T and Crook J (2007). Modelling and predicting loss given default for credit cards. In: *Credit Scoring and Credit Control XI conference*.

Bellotti T and Crook J (2009). Macroeconomic conditions in models of Loss Given Default for retail credit, *Credit Scoring and Credit Control XI Conference*, August

Benjamin N, Cathcart A and Ryan K (2006). Low Default Portfolios: A Proposal for Conservative Estimation of Default Probabilities. *Discussion Paper*, *Financial Services Authority*.

Benzschawel T, Haroon A and Wu T (2011). A Model for Recovery Value in Default, *Journal of Fixed Income*, **21**(2), 15-29.

Berger J (1980). Statistical Decision Theory: Foundations, Concepts and Methods, Springer-Verlag.

Berger J and Berliner L (1986). Robust bayes and empirical bayes analysis with contaminated priors, *The Annals of Statistics*, **14**(2), 461–486.

Berry M and Linoff S (2000). Mastering data mining: The art and science of customer relationship management, John Wiley & Sons, Inc, New York.

Bi J and Bennet KP (2003). Regression error characteristic curves. In: *Twentieth International Conference on Machine Learning*, Washington DC, USA.

Bishop CM (1995). Neural Networks for Pattern Recognition. Oxford University Press: Oxford, UK.

Bonfim D (2009). Credit risk drivers: Evaluating the contribution of firm level information and of macroeconomic dynamics, *Journal of Banking & Finance*, **33**(2), 281-299.

Box G and Cox D (1964). An analysis of transformations. *Journal of Royal Statistics Society*, **26**, 211-252.

Box G and Tiao G (1992), Bayesian Inference in Statistical Analysis, John Wiley & Sons, New York.

Breiman L (2001). Random Forests. *Machine Learning*, **45**(1), 5-32.

Breiman L, Friedman J, Stone C, and Olshen R (1984). *Classification and Regression Trees*. Chapman & Hall/CRC.

Burgt M (2007). Calibrating Low-Default Portfolios, using the Cumulative Accuracy Profile, ABN AMRO Group Risk Management Tools and Modelling.

Carling K, Jacobson T, Lindé J and Roszbach K (2007). Corporate Credit Risk Modelling and the Macroeconomy, *Journal of Banking & Finance*, **31**(3), 845-868.

Caselli S and Querci F (2009). The sensitivity of the loss given default rate to systematic risk: New empirical evidence on bank loans. *Journal of Financial Services Research*, **34**, 1-34.

Cespedes JCG, de Juan Herrero JA, Rosen D, and Saunders D (2010). Effective Modeling of Wrong Way Risk, Counterparty Credit Risk Capital and Alpha in Basel II, *Journal of Risk Model Validation*, **4**(1), 71-98.

Chaloner KM and Duncan GT (1983). Assessment of a beta prior distribution: Pm elicitation, *The Statistician*, **32***(1/2, Proceedings of the 1982 I.O.S. Annual Conference on Practical Bayesian Statistics)*, 174–180.

Chalupka R and Kopecsni J (2009). Modeling Bank Loan LGD of Corporate and SME Segments: A Case Study, *Czech Journal of Economics and Finance*, **59**(4), 360-382

Chatterjee S and Barcun S (1970). A Nonparametric Approach to Credit Screening. *Journal of the American Statistical Association*, **65**(329), 50-154.

Chatterjee S, Corbae DP, Nakajima M and Rios-Rull JV (2007). A Quantitative Theory of Unsecured Consumer Credit with Risk of Default, *Econometrica*, **75**(6), 1525-1589.

Chawla NW, Bowyer KW, Hall LO and Kegelmeyer WP (2002). SMOTE: Synthetic Minority Over-sampling Technique. *Journal of Artificial Intelligence Research*, **16**, 321–357.

Chou M (2006). Cash and credit card crisis in Taiwan, *Business Weekly*, 24–27.

Cohen P, Cohen J, West S and Aiken L (2002). Applied multiple regression/ correlation analysis for the behavioral sciences. Lawrence Erlbaum

Crouhy M, Galai D and Mark R (2000). A comparative analysis of current credit risk models, *Journal of Banking & Finance*, **24**, 57–117.

Davis RH, Edelman DB and Gammerman AJ (1992). Machine learning algorithms for credit card applications, *IMA Journal of Mathematics Applied Business & Industry*, **4**, 43-51.

DeGroot MH (1970). Optimal Statistical Decisions, McGraw-Hill.

DeLong ER, DeLong DM, and Clarke-Pearson DL (1988). Comparing the Areas Under Two or More Correlated Receiver Operating Characteristic Curves: A Nonparametric Approach. *Biometrics*, **44**(3), 837–845.

Demšar J (2006). Statistical Comparisons of Classifiers over Multiple Data Sets. *Journal of Machine Learning Research*, **7**, 1-30.

Desai VS, Crook JN and Overstreet Jr GA (1996). A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operation Research Society*, **95**(1), 24-37.

Diaconis P and Ylvisaker D (1985). Quantifying Prior Opinion, In: *Bayesian Statistics* vol. 2 (In: Bernardo JM, DeGroot MH, Lindley DV and Smith AFM, Editors), Elsevier Science Publishers BV, North-Holland, 133–156.

Draper N, and Smith H (1998). Applied Regression Analysis. Wiley.

Duffie D and Singleton KJ (1998). Simulating correlation defaults, Bank of England Conference on Credit Risk Modeling and Regulatory Implications, London, 21–22 September.

Dwyer DW (2006). The distribution of defaults and bayesian model validation, *Technical report, Moody's/KMV*, November.

Erdem C (2008). Factors affecting the probability of credit card default and the intention of card use in Turkey. *International Research Journal of Finance and Economics*, **18**, 1450-2887.

Fawcett T (2006). An introduction to roc analysis. *Pattern Recognition Letters*, **27**, 861-874.

Fernandes JE (2005). Corporate credit risk modeling: Quantitative rating system and probability of default estimation, mimeo.

Financial Supervision Authority, UK (2004a). Issues arising from policy visits on exposure at default in large corporate and mid market portfolios. Working Paper, September.

Financial Supervision Authority, UK (2004b). Own estimates of exposure at default. Working Paper, November.

Crone S and Finlay S (2011). Instance Sampling in Credit Scoring: an empirical study of sample size and balancing, *International Journal of Forecasting*, forthcoming; Originally in: Big or Balanced? An empirical study of the effects of sample size and balancing on model performance, *In: Conference on Risk Management in the Personal Financial Services Sector*, 22-23 Jan 2009, Imperial College, London

Fogarty TC, Ireson NS, and Battles SA (1992). Developing rule based systems for credit card applications from data with genetic algorithms. *IMA Journal of Mathematics Applied In Business and Industry*, **4**, 53-59.

Forrest A (2005). Likelihood Approaches to Low Default Portfolios, Joint Industry Working Group Discussion Paper

Freed N and Glover F (1981). Simple but powerful goal programming models for discriminant problems. *European Journal of Operational Research*, **7**, 44-60.

Freund R and Littell R (2000). SAS System for Regression. Wiley.

Friedman J (2001). Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, **29**(5), 1189-1232.

Friedman J (2002). Stochastic gradient boosting. *Computational Statistics & Data Analysis*, **38**(4), 367-378.

Friedman JF (1991). Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1-141.

Friedman M (1940). A comparison of alternative tests of significance for the problem of m rankings. *The Annals of Mathematical Statistics*, **11**(1), 86-92.

Garthwaite PH et al. (2005). Statistical methods for eliciting probability distributions, *Journal of the American Statistical Association,* **100**, 680–701.

Giambona F, and Iancono VL (2008). Survival models and credit scoring: some evidence from Italian Banking System., 8th international business research conference, Dubai, 27th-28th March 2008

Giudici P (2001). Bayesian data mining, with application to benchmarking and credit scoring, *Applied Stochastic Models in Business and Society,* **17**, 69–81.

Giudici P (2003). Applied data mining: Statistical methods for business and industry, John Wiley & Sons, Inc, New York.

Gordy MB (2000). A comparative anatomy of credit risk models, *Journal of Banking & Finance* **24**, 119–149.

Gruber W, and Parchert R (2006). Overview of EAD estimation concepts, in: Engelmann B, Rauhmeier R (Eds), *The Basel II Risk Parameters: Estimation, Validation and Stress Testing,* Springer, Berlin, 177-196.

Grunert J and Weber M (2008). Recovery rates of commercial lending: Empirical evidence for German companies. *Journal of Banking & Finance*, **33**(3), 505–513.

Guettler A and Liedtke HG (2007). Calibration of Internal Rating Systems: The Case of Dependent Default Events, *Kredit und Kapital*, **40**, 527-552

Gupton G and Stein M (2002). Losscalc: Model for predicting loss given default (lgd). Tech. rep., Moody's.

Gupton G and Stein M (2005) LossCalc V2: Dynamic prediction of LGD. Moody's Investors Service.

Hampel F, Ronchetti R, Rousseeuw P and Stahel W (1986). Robust statistics: the approach based on influence functions. Wiley.

Han J and Kamber M (2001). Data mining: Concepts and techniques, Morgan Kaufmann, San Fransisco.

Hand DJ and Henley WE (1997). Statistical classification methods in consumer credit scoring: A review, *Journal of the Royal Statistical Society, Series A – Statistics in Society*, **160**(3), 523–541.

Hand DJ and Jacka SD (1981). Discrimination and Classification. Wiley.

Hanson S and Schuermann T (2006). Confidence Intervals for Probabilities of Default, *Journal of Banking & Finance*, **30**(8), 2281-2301.

Hartmann-Wendels T and Honal M (2006). Do economic downturns have an impact on the loss given default of mobile lease contracts? An empirical study for the German leasing market. Working Paper, University of Cologne

Hastie T, Tibshirani R and Friedman J (2001). The Elements of Statistical Learning, Data Mining, Inference, and Prediction. Springer: New York.

Henley WE and Hand DJ (1997). Construction of a *k*-nearest neighbour credit-scoring system. *IMA Journal of Management Mathematics*, **8**(4), 305-321.

Hlawatsch S and Ostrowski S (2010). Simulation and Estimation of Loss Given Default, *FEMM Working Papers 100010*, Otto-von-Guericke University Magdeburg, Faculty of Economics and Management.

Hlawatsch S and Reichling P (2010). A Framework for LGD Validation of Retail Portfolios, *Journal of Risk Model Validation*, **4**(1), 23-48

Hoerl AE and Kennard RW (1970). Ridge regression: Biased estimation for nonorthogonal problems. *Technometrics*, **12**, 55-67.

Holland P and Welsch R (1977). Robust regression using iteratively reweighted least squares. *Communications in Statistics: Theory and Methods*, **6**, 813-827.

Hosmer DW and Stanley L (2000). Applied Logistic Regression, 2nd ed. New York; Chichester, Wiley.

Hu YT and Perraudin W (2002). The dependence of recovery rates and defaults, Mimeo, Birkbeck College.

Huang CL, Chen MC and Weng CJ (2007). Credit scoring with a data mining approach based on support vector machines, *Expert Systems with Applications*, **33**(4), 847-856.

Huber P (1964). Robust estimation of a location parameter. *Annals of Mathematical Statistics*, **35**, 73-101.

Huber P and Ronchetti E (2009). Robust statistics. Wiley.

Jacobs M (2008). An Empirical Study of Exposure at Default. OCC Working Paper. Washington, DC: Office of the Comptroller of the Currency.

Jacobs M and Karagozoglu AK (2011). Modeling Ultimate Loss Given Default on Corporate Debt, *Journal of Fixed Income*, **21**(1), 6-20.

Jagielska I and Jaworski J (1996). Neural network for predicting the performance of credit card accounts, *Computational Economics*, **9**(1), 77–82.

Jain AK et al. (2000). Statistical pattern recognition: A review, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **22**(1), 4–37.

Japkowicz N (2000). Learning from imbalanced data sets: a comparison of various strategies. *AAAI Workshop on Learning from Imbalanced Data Sets*, **6**, 10-15.

Jaynes ET (2003). Probability Theory: the Logic of Science, Cambridge University Press, New York.

Jiménez G, Lopez J A, and Saurina J (2009). EAD Calibration for Corporate Credit Lines. *Journal of Risk Management at Financial Institutions*, **2**, 121–29.

Kadane JB and Wolfson LJ (1998). Experiences in elicitation, *The Statistician*, **47**(1), 3–19.

Kadane JB, Dickey JM, Winkler RL, Smith WS and Peters SC (1980). Interactive elicitation of opinion for a normal linear model, *Journal of the American Statistical Association* **75**(372), 845–854 Dec.

Kahneman D and Tversky A (1974). Judgement under uncertainty: heuristics and biases, *Science*, **185**, 1124–1131.

Kennedy K, Mac Namee B, Delany SJ (2011). Using semi-supervised classifiers for credit scoring, *The Journal of the Operational Research Society*, to appear.

Kiefer NM (2010). Default Estimation and Expert Information. *Journal of Business and Economic Statistics*, **28**(2), 320-328.

Koh HC and Chan KLG (2002). Data mining and customer relationship marketing in the banking industry, *Singapore Management Review*, **24**(2), 1–27.

Kohavi R (1995). Wrappers for performance enhancement and oblivious decision graphs. PhD thesis, Department of Computer Science, Stanford University.

Kolesar P and Showers JL (1985). A robust credit screening model using categorical data. *Management Science*, **31**, 123-133.

Lee YS et al. (2002). Credit scoring using the hybrid neural discriminant technique, *Expert Systems with Applications*, **23**(3), 245–254.

Lee YS et al. (2004). A data mining approach to constructing probability of default scoring model. In *Proceedings of 10th conference on information management and implementation*, 1799–1813.

Lessmann S, Baesens B, Mues C, and Pietsch S (2008). Benchmarking classification models for software defect prediction: A proposed framework and novel findings. *IEEE Transactions on Software Engineering*, **34**, 485-496.

Lewis EM (1990). An Introduction to Credit Scoring, Athena Press, San Rafael.

Li H (2010). Downturn LGD: A Spot Recovery Approach, *MPRA Paper 20010*, University Library of Munich, Germany.

Lindley DV (1982). The improvement of probability judgements, *Journal of the Royal Statistical Society. Series A (General)* **145**(1), 117–126.

Lindley DV, Tversky A and Brown RV (1979). On the reconciliation of probability assessments, *Journal of the Royal Statistical Society. Series A (General)* **142**(2), 146–180.

Loterman G, Brown I, Martens D, Mues C, and Baesens B (2009). Benchmarking State-of-the-Art Regression Algorithms for Loss Given Default Modelling *11th Credit Scoring and Credit Control Conference (CSCC XI)*. Edinburgh, UK

Luo X and Shevchenko PV (2010). LGD credit risk model: estimation of capital with parameter uncertainty using MCMC, *Quantitative Finance Papers*

Martin D (1977). Early warning of bank failure: A logit regression approach. *Journal of Banking and Finance*, **1**, 249–276.

Martens D, Baesens B, and Van Gestel T (2009). Decompositional rule extraction from support vector machines by active learning. *IEEE Transactions on Knowledge and Data Engineering*, **21**, 178-191.

Martens D, Baesens B, Van Gestel T and Vanthienen J (2007). Comprehensible credit scoring models using rule extraction from support vector machines. *European Journal of Operational Research*, **183**, 1466-1476.

Matuszyk A, Thomas L C and Mues C (2010). Modelling LGD for Unsecured Personal Loans: Decision Tree Approach. *Journal of Operational Research Society*, **61**(3), 393-398.

Mays E (1998). Credit Risk Modeling: Design and Application, New York: Glenlake

Merton RC (1974). On the Pricing of Corporate Debt: The Risk Structure of Interest Rates, *Journal of Finance*, 29(2), 449–470.

Mester LJ (1997). What's the Point of Credit Scoring? *Business Review*, **5**, 3-16.

Miu P and Ozdemir B (2006). Basel requirements of downturn loss given default modelling and estimating probability of default and loss given default correlations. *The Journal of Credit Risk*, **2**(2), 43-68

Miyake M and Inoue H (2009). A Default Probability Estimation Model: An Application to Japanese Companies. *Journal of Uncertain Systems*, **3**(3), 210–220

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

Murphy A, Kocagil A, Escott P and Glormann F (2002). Moody's RiscCalc for private companies: Portugal. *Moody's Investors Service Rating Methodology*.

Nagelkerke NJD (1991). A note on a general definition of the coefficient of determination. *Biometrica*, **3**, 691–692.

Nelson B et al. (2003). An error rate comparison of classification methods with continuous explanatory variables, *IIE Transactions*, **35**, 557–566.

Nemenyi PB (1963). Distribution-free multiple comparisons. PhD thesis, Princeton University.

OCC (2006). Validation of credit rating and scoring models: a workshop for managers and practitioners, *Office of the Comptroller of the Currency* (2006).

Ohlson J (1980). Financial ratios and the probabilistic prediction of bankruptcy. *Journal of Accounting Research*, 109–131.

Pluto K and Tasche D (2005). Thinking positively, *Risk*, 72–78 August.

Provost F, Jensen D, and Oates T (1999). Efficient progressive sampling. *In Proceedings of the Fifth International Conference on Knowledge Discovery and Data Mining*. ACM Press.

Qi M (2009). Exposure at Default of Unsecured Credit Cards, Credit Risk Analysis Division, Washington, DC: Office of the Comptroller of the Currency.

Qi M and Zhao X (2011). Comparison of Modeling Methods for Loss Given Default, *Journal of Banking & Finance*, **35**(11), 2842-2855.

Quinlan JR (1993). C4.5 Programs for Machine Learning. Morgan Kaufmann: San Mateo, CA.

Raiffa H and Schlaifer R (1961). Applied Statistical Decision Theory, Harvard Business School.

Rosenberg E and Gleit A (1994). Quantitative methods in credit management: A survey, *Operations Research* **42** (4), 589–613.

SAS Institute (2002) .Comply and Exceed. Credit Risk Management for Basel II and Beyond. A SAS White Paper.

Schuermann T (2004). What do we know about loss given default? Working Paper No. 04-01, Wharton Financial Institutions Center, Feb.

Shleifer A and Vishny R (1992). Liquidation values and debt capacity: A market equilibrium approach, *Journal of Finance*, **47**, 1343-1366.

Sigrist F and Stahel WA (2010). Using The Censored Gamma Distribution for Modeling Fractional Response Variables with an Application to Loss Given Default, *Quantitative Finance Papers*

Smithson M and Verkuilen J (2006). A better lemon squeezer? maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, **11**, 54-71.

Somers M and Whittaker J (2007). Quantile regression for modelling distribution of profit and loss. *European Journal of Operational Research*, **183**(3). 1477-1487

Steenackers A and Goovaerts MJ (1989). A credit scoring model for personal loans. *Inurances: Mathematics and Economics*, **8**(1), 31-34.

Sufi A (2009). Bank Lines of Credit in Corporate Finance: An Empirical Analysis. *The Review of Financial Studies*, **22**(3), 1057-1088.

Suykens JAK, Van Gestel T, De Brabanter J, De Moor B and Vandewalle J (2002). Least Squares Support Vector Machines, World Scientific, Singapore.

Taplin R, Huong M and Hee J (2007). Modeling exposure at default, credit conversion factors and the Basel II Accord. *Journal of Credit Risk*, **3**(2), 75-84.

Tarashev NA (2008). An Empirical Evaluation of Structural Credit Risk Models *International Journal of Central Banking*, **4**(1), 1-53.

Tasche D (2003). A Traffic Lights Approach to PD Validation, Frankfurt.

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

Valvonis V (2008). Estimating EAD for retail exposures for Basel II purposes. *Journal of Credit Risk*, **4**(1), 79-109

Van Der Burgt M (2007). Calibrating Low-Default Portfolios, using the Cumulative Accuracy Profile. ABN AMRO.

Van Gestel T and Baesens B (2009). *Credit Risk Management*. Oxford University Press.

Van Gestel T, Baesens B, Van Dijcke P, Garcia J, Suykens J and Vanthienen J (2006). A process model to develop an internal rating system: Sovereign credit ratings. *Decision Support Systems*, **2**, 1131-1151.

Van Gestel T, Baesens B, Van Dijcke P, Suykens J, Garcia J, and Alderweireld T (2005). Linear and non-linear credit scoring by combining logistic regression and support vector machines. *Journal of Credit Risk*, **1**.

Van Gestel T, Martens D, Feremans D, Baesens B, Huysmans J and Vanthienen J (2007). Forecasting and analyzing insurance companies' ratings. *International Journal of Forecasting*, **23**, 513-529.

Van Gestel T, Suykens J, Baesens B, Viaene S, Vanthienen J, Dedene G, De Moor B, and Vandewalle J (2003). Benchmarking least squares support vector machine classifiers. *Machine Learning*, **54**, 5-32.

Vapnik V (1995). The Nature of Statistical Learning Theory. Springer.

Viganò L (1993). "A Credit Scoring Model for Development Banks: An African Case Study", *Savings and Development*, **17**(4), 441-482.

Walker SH and Duncan DB (1967). Estimation of the Probability of an Event as a Function of Several Independent Variables. *Biometrika*, **54**, 167-179

Wang H and Hu D (2005). Comparison of SVM and LS-SVM for regression. *International Conference on Neural Networks and Brain*, **1**, 279-283.

Weiss GM and Provost FJ (2003). Learning When Training Data are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research (JAIR)*, **19**, 315-354.

West D (2000). Neural network credit scoring models. *Computers & Operational Research*, **27**(11-12), 1131–1152.

Wiginton JC (1980). A Note on the Comparison of Logit and Discriminant Models of Consumer Credit Behavior. *Journal of Financial and Quantitative Analysis*, **15**, 757-770.

Wilde T and Jackson L (2006). Low-default portfolios without simulation, *Risk*, 60–63.

Witten IH and Frank E (2005). Data Mining: Practical machine learning tools and techniques, 2nd Edition, Morgan Kaufmann, San Francisco.

Yang Y (2007). Adaptive credit scoring with kernel learning methods. *The European Journal of Operational Research Society*, **183**(3), 1521-1536.

Yao P (2009). Hybrid Classifier Using Neighborhood Rough Set and SVM for Credit Scoring, *International Conference on Business Intelligence and Financial Engineering*, 138-142

Yeh IC and Lien CH (2009). The comparisons of data mining techniques for the predictive accuracy of probability of default of credit card clients, *Expert Systems with Applications*, **36**(2), 2473-2480

Yobas MB, Crook JN and Ross P (2000). Credit scoring using neural and evolutionary techniques. *IMA Journal of Management Mathematics*, **11**(2), 111-125.

Zellner A (1996). Introduction to Bayesian Inference in Econometrics, John Wiley & Sons, New York.