Tackling change and uncertainty in credit scoring

Thesis

Tackling change and uncertainty in credit scoring

Mark Gerard Kelly, BSc, MSc.

Department of Statistics
Faculty of Mathematics and Computing
The Open University

Submitted for the degree of Doctor of Philosophy
October 1998

Abstract

Credit scoring methods summarise information on credit applicants. An assessment of creditworthiness is derived from this summary. This thesis is concerned with statistical methods of credit scoring.

Much of the existing literature on credit scoring is concerned with comparing the predictive power of a wide variety of classification techniques. However, much of the published work concludes that classifier performance on credit data is relatively insensitive to the choice of statistical technique. Consequently, the techniques used in commercial credit scoring have remained broadly similar during recent years. This thesis investigates credit scoring from a more fundamental level, by considering the formulation of the credit problem.

A review of the credit literature is given, focusing on areas that have been subjected to much recent research activity. Details of the data sets used throughout this thesis are provided and analysed using techniques common to the credit industry.

Methods that capitalise on the uncertainty and flexibility in the definitions of the classes used to represent 'good' and 'bad' credit risks are proposed. Firstly, a class of models is described that permits the choice of class definition to be deferred until the time at which the classification is required. Secondly, a strategy for choosing a suitable definition which optimises some external criterion is introduced. In addition, an approach is presented that combats classifier deterioration resulting from the evolution of the underlying populations.

This thesis is essentially concerned with the uncertainties and change inherent in credit scoring. We present novel ways in which these properties may be incorporated in the formulation of the credit problem.

Acknowledgements

I would like to thank my supervisor, Professor David Hand, for his constant guidance and encouragement. I would also like to thank Dr Niall Adams for his help and support.

I am grateful to Dr Sam Korman from Abbey National plc for his interest in this project and for supplying several financial data sets used in this thesis. Additional data were made available by Dr Andrew Webb from the Defence Evaluation and Research Agency.

This project was funded by a CASE award from the Engineering and Physical Sciences Research Council with additional funds from Abbey National plc.

Contents

Chapter 1

Introduction

In recent years ideas from statistics and pattern recognition have become an integral part of the collection of techniques used for credit scoring purposes. The consumer credit industry uses these techniques to assess the creditworthiness of applications for credit products. Ultimately the goal of a financial institution is to maximise overall profit while also minimising losses due to customers who do not adhere to their credit agreement. Recent PhDs (Henley 1995, Sewart 1997) have taken a technique oriented approach. Specific problems inherent in credit scoring have been addressed as well as the evaluation of techniques previously omitted from the credit scoring literature. More generally, the credit literature includes many comparative studies of different classification techniques. Thomas (1998) noted that credit scoring performance is surprisingly insensitive to the actual statistical technique employed to make classifications. In the hope of gaining deeper insight into the mechanics of the problem, we investigate issues related to the way in which the problem is formulated rather than investigating more complex classification methods.

In the next section we describe the ideas underpinning credit scoring and briefly comment on the rise of the industry during the twentieth century. Scorecards, a special implementation of classification rules, are central to the application of scoring. We use the terms scorecard and classifier interchangeably throughout this thesis. Section 1.2 provides a review of the credit scoring literature. Particular emphasis is given to areas which have provided popular research topics in recent years, such as behavioural scoring and profitability modelling. The

literature review also suggests that no one technique has been more successful than any other. Finally, Section 1.3 introduces the various classifier assessment criteria in common use throughout the industry and discusses their shortcomings.

In later chapters we use financial data sets extensively. These data sets are described in Chapter 2. We use two sets of data. First, unsecured personal loan data that consists of information on the account holder as well as the manner in which the loan was conducted. Second, data describing the monthly transactions and conduct of current accounts. Features not commonly found in other data sets are described along with data collection and sampling strategies used to gather data. Finally, in Chapter 2 we give examples of problems encountered when dealing with large credit data sets, which, if unnoticed, may substantially influence results.

Chapter 3 outlines common practices frequently encountered in credit scoring. We illustrate the process of scorecard construction using techniques common to the industry – linear and logistic regression. The results obtained are comparable with those that would be obtained in a commercial environment. These results are also consistent with reports in the literature.

Much of this thesis is based on the notion of uncertainty and ambiguity in the class definitions used in credit applications. For example a typical definition of bad may be represented by the magnitude of arrears on an unsecured personal loan, exceeding some limit. However, the precise choice of this limit may be subject to uncertainty. We assert that definitive definitions of good and bad classes do not exist. This is signified by the diversity of the class definitions used in the credit industry. Given that different definitions may be acceptable, coupled with the willingness of credit practitioners to experiment with other definitions, we aim to construct models that are resistant to changing definitions as well as models that maximise performance by capitalising on flexibility within the definitions.

Two different approaches to credit scoring are investigated in Chapter 4. Section 4.2 investigates the technique of survival analysis. We use survival techniques to predict a customers' likely survival time on the basis of their feature vector. Section 4.3 investigates the effect of different class definitions on the predictive power of the scorecards. Alternative outcomes such as early loan settlement are investigated. In addition we explore the common practice of collectively treating products with different characteristics in the same way. We demonstrate that improvements can be made by considering the properties of each product.

In practice when a scorecard is built it uses a specific definition. However, due to external influences, such as the economic climate, it may prove beneficial to change the definition used. Consequently a model that embodies many alternative definitions may be useful, such global models are proposed in Chapter 5. These models permit a change in class definition without the need to explicitly re-estimate the parameters of the model. We compare the performance of global models with that of a standard model under a range of situations in which a change of definition may have been necessary.

Class definitions used for current accounts rely on several variables. The value taken by each of these variables is compared to a threshold and the results combined to determine the definition. The values taken by each of these definition variables may be varied between specified limits. In Chapter 6 we propose a mechanism that can be used for searching a large group of plausible definitions so that a choice can be made which maximises some external performance criterion. We find that equally acceptable definitions can result in much improved classification performance. In addition, we suggest methods that can be used to compare the relative performance of the different definitions.

Finally, in Chapter 7 we tackle problems that originate from external sources. Methods are introduced to combat *population drift* – the tendency of populations to evolve with time. We show that the variables used in scorecard construction are subject to drift. Furthermore, we demonstrate that these drifting variables

result in deteriorating scorecard performance. An adaptive approach is suggested which permits new observations to be incorporated into the scorecard without the need to completely re-estimate the model.

In Chapter 8 we summarise the work presented in this thesis. We suggest that incorporation of uncertainties, such as those presented in this thesis, into the credit scoring problem is most likely to result in sustainable improvements in classification performance. Finally we suggest areas where further research may be fruitful.

1.1 Introduction to Credit Scoring

In recent years credit has become part of most of our lives, whether it be through a mortgage, credit card, unsecured personal loan, hire purchase agreement or overdraft. The number of credit providers has also been rising rapidly, resulting in an extremely competitive credit market place. Competition among credit providers has led to better deals for the consumer. More favourable interest rates and incentives are offered to encourage customers to do business. This increased competition with reduced profit margins has made it even more important that credit providers grant credit only to applicants who are thought to have low risk of defaulting on the agreements. That is, those who are likely to repay the amount borrowed in accordance with the credit agreement. Substantial efforts have been made to design techniques that minimise bad debt by assessing the risk associated with credit applications. These techniques are often referred to as *credit scoring*.

1.1.1 The Emergence of Credit Scoring

Lewis (1992a) discusses credit transactions dating back thousands of years. However these distant roots have little relevance to the multibillion pound credit industry of today. Lewis amongst others also states that the introduction of the car at the turn of the twentieth century had a major influence on credit. Many people wished to become car owners but were unable to pay the full amount in

advance. Originally banks were reluctant to supply credit for this purpose. However, independent finance companies were founded to meet public demand and these companies became successful and profitable, resulting in banks changing their policies regarding such matters. The banking community quickly re-established their dominance in all areas of credit provision. From this point on, the credit industry grew at a moderate pace with product specific loans growing in popularity and the introduction of mail order proving to be a successful marketing innovation.

The 1960's saw the introduction of the credit card but it was not until the 1970's that the product established itself in the credit market with the advent of Visa and Mastercard – these products are still market leaders today. This product has been the most significant innovation in the credit industry to date (Lewis, 1992a). A credit card holder was able to make purchases at any goods or service outlet which had agreed to accept credit cards as a method of payment. This new product ensured the continuing swift growth of the credit business. This was a new type of credit, one which allowed the holder to buy any product or service at any time, without the need for a consultation, form filling or pre-agreement from any third party. The cost of any purchases made could then be repaid at a later date at whatever repayment scheme, within certain restrictions, the credit card holder preferred. In consequence credit cards quickly became a popular product amongst the general public. In 1972 Mastercard, was introduced followed by Visa in 1977. A rapid rise has since taken place both in the number of cards in circulation and spending on those accounts. In 1980, 11.7 million credit cards were in circulation in the UK, spending on these cards totalled £2.9 billion. During the years since 1980 the economy has been in recession twice. However, in 1997 the number of credit cards held in the UK had more than tripled, 37.1 million cards with spending at £62.7 billion – an increase by a factor of twenty.

During the developmental phase the credit granting decision was very much judged on merit – judgmental analysis. In the early days of credit each case was judged subjectively, 'You can't make a good credit decision unless you can see

the white of the customer's eyes', was an attitude commonplace twenty years ago, (Coe, 1997). As the credit industry grew this need for a personal seal of approval for each application began to cause problems itself. The processing time required for such a large volume of applications resulted in a considerable lag between application and final decision. As the credit marketplace was becoming more competitive a backlog such as this could mean lost custom. The consumer may no longer tolerate long delays in processing their application, but instead seek credit from a competitor.

The credit industry was desperately in need of an automated decision process to reduce the time needed to deal with the huge volume of new applications. Score tables were developed which generalised each application question into categories. These categories would have points or scores associated which could be assigned to customers so that application decisions could be made on the basis of total score. Durand (1941) gave the first published account of a scoring mechanism based on discriminant analysis and applied to finance. However, credit providers did not readily adopt scoring systems and it was not until much later such methods were implemented in the consumer area.

During the 1970's and 1980's when the credit market was quickly expanding, tremendous advances in the power of computers and data storage were taking place. Statistical methodology coupled with technological advances provided the necessary tools to develop automated scoring systems capable of dealing with the increased demand for credit.

1.1.2 What is Credit Scoring?

Most people are familiar with the tedious forms that must be completed when applying for most types of credit. The completed form is then 'processed' while the applicant awaits a decision on their application. Credit scoring takes place during this application processing phase.

Credit scoring is a means of summarising information on an application to produce a number called a *score.* This score is then compared with a predetermined threshold. Crudely, if the applicant's score is greater than the threshold then credit is granted, otherwise credit is declined. In practice the decision is not quite so clear cut. A grey area surrounds the threshold, concerned with the question "why should applications scoring two or three points above the threshold be granted credit, but those scoring just below the threshold be declined?" Applications whose scores fall in this area may be referred to a further decision stage. Referral after an inconclusive credit score often returns the final decision to an individual.

In comparison to judgmental analysis, statistical methods for obtaining scores should produce more consistent predictions of future behaviour, utilise a precise definition of creditworthiness and produce a solution which, dealing with many applications, would result in a lower cost per application than judgmental analysis. Such systems were based on the notion that applicants with similar traits would be expected to exhibit similar patterns of repayment. Scoring systems were quickly preferred over judgmental analysis in the decision making process. A decision on granting credit could be related to the perceived risk associated with the application rather than some poorly defined notion of a poor risk assessed purely by the judgement of a human agent employed by the credit provider. Chandler and Coffman (1979) and Wilkinson (1992) have contrasted the two approaches based on automated systems and judgement. The ideas of credit scoring have often been criticised, see Johnson (1992) and Capon (1982). A common criticism is that credit scoring assesses the risk associated with sub-populations rather than individuals. Nonetheless, credit scoring has become the credit industry's standard approach to assessing new credit applications.

1.1.3 Scorecards

A scorecard is the tool used in the credit industry to assess applications for credit products. A scorecard comprises a table of variables that have been pre-selected.

Each variable is divided into a number of levels. Each level of each variable has an associated number or score. When a new applicant arrives in the system their details are categorised and compared to the scorecard. The applicant's credit score is the sum of the scores for each variable. This final score is then compared with a threshold to determine whether the application is perceived to be creditworthy.

More formally a scorecard is based on some statistical model which generates an estimate of the probability that an application will default at some stage during its lifetime. These probability estimates are used to select which applications should be granted credit.

A scorecard is equivalent to the statistical model from which it is derived. Throughout this thesis we will refer to estimated probabilities and scores as well as models and scorecards. These terms are analogous and so will be used interchangeably.

1.1.4 Data

Enormous quantities of data are usually available for modelling financial applications. Data used in scorecard construction can fall into one of three types, detailed below. Specifics of data sets that will be used frequently in this thesis can be found in Chapter 2.

Application Data

Data of this type comes directly from completed application forms. Applicants' details of income, financial commitments, job details, residential status as well as information on current credit agreements are often included.

Bureau Data

In England there exist several credit reference agencies. These agencies hold information on the majority of the population who are registered on the electoral roll. Financial service outlets such as banks, building societies and insurance

firms may purchase this information. Information available will typically consist of credit irregularities at addresses associated with the applicant, county court judgements against individuals and the number of recent applications for credit made by the applicant. There is also *credit account information sharing*, which is data connected with outstanding credit agreements, which is shared via the credit reference agencies, between financial institutions.

Bureau data is usually very predictive, that is, it is useful for making correct predictions of applicants' risk. Some companies may use such data to enforce *policy declines.* If a particular factor, such as a county court judgement, is present on an application, that application may be rejected regardless of any other information.

Behavioural Data

Behavioural data is information collected during the operational lifetime of a credit product that an existing customer holds. For example, repayments on a loan or credit card, or debits and credits of a current account. These data are used for account monitoring. This is known as behavioural scoring and is discussed in Section 1.2.2.

1.1.5 Issues and Problems with Scorecard Construction

Superficially, the construction of a scorecard is a very simple process. However, there are many intricacies that should not be overlooked. These intricacies can easily lead to inaccurate or misleading results. Crucial elements when building a scorecard are data, statistical understanding, knowledge of the financial area and computing power. Several interesting points arise and are detailed below.

1.1.5.1 Legal Issues

Legislation has sought to eliminate prejudice from credit scoring. The Consumer Credit Act, Watson (1995) and the Data Protection Act have ensured that unfair discrimination based on sex, race and religion cannot take place. Hand and

Henley (1997a) suggests that it may be more appropriate to include all risk predictors. Prohibiting the use of certain predictor variables inevitably leads to some groups being assigned higher default probabilities than their true value. Hsia (1978) discusses the implications of the Equal Opportunities Act to credit scoring. Inclusion of predictor variables related to such information is outlawed. Johnson (1992) discusses these issues, and others relevant to the US where strict policies are enforced, even with regard to age.

1.1.5.2 Data Manipulation

With ample amounts of data, the data set is randomly divided into two parts, the design and test sets. This is a common procedure in credit scoring. However, credit data is often collated over time, by randomly splitting the data set the possibility of detecting patterns that evolve with time (see Chapter 7) is lost. The design set is used to estimate the classification rule and the test set used to obtain an independent estimate of the classifier's performance. The proportions of this split may vary depending on company policy, the amount of available data or the individual employed to construct the scorecard. Commonly splits between 50%-50% and 80%-20% are used. The test sample is sometimes referred to as a *holdout sample* in the literature.

There is a large literature on re-sampling and simulation methods, which are techniques that may be useful when ample data are not available. However, insufficient data is not a problem frequently encountered in credit scoring.

1.1.5.3 Biasing of Results

The performance of a model, constructed on the design set, is often assessed using the test set. In the credit industry this performance measure is often used as a further measure of classifier performance. If test set performance is inadequate then models may be changed in an attempt to enhance model performance. This results in a model that is tuned to the test sample. Consequently, it is unlikely that observed improvements would be preserved in future independent samples

introduced to the scorecard. This and other types of overfitting are discussed further in Section 1.1.6.

1.1.5.4 Missing Values

Missing values are present in most large data sets. There are a variety of causes of missing data, including storage system errors, incorrectly keyed applications or *soft fraud.* Applicants who 'overlook' certain application questions in the belief that omitting to answer a question will improve their chance of a successful application are committing soft fraud.

Missing values are often entered into the data set as a category. The justification for this is that missing values on application forms may contain information about creditworthiness. Entering missing values in this way seems a reasonable thing to do, as any predictive power demonstrated by the missing values may be incorporated into the model. However, by grouping all missing values together we are unable to distinguish between the suspicious fraudulent missing values which may contribute towards the predictive power of the scorecard and erroneous missing values which are presumably random, or at least independent of any mechanism that can be used to predict future behaviour.

1.1.5.5 Scorecard Splitting

Development of scorecards often involves *splitting.* This means that for a particular variable a different scorecard is built for different levels of that variable. For example, age may be categorised into four levels. Under these conditions four indicator variables corresponding to the four levels may be used in scorecard construction. Constructing a scorecard split by age would produce four different scorecards, one for each of the sub populations defined according to each level of age. This splitting technique effectively involves an interaction term between the splitting variable and all other predictors in the model. It was mentioned in Section 1.1.5.1 that the use of certain variables such as sex is not permitted. Chandler and Ewert (1976) examine the effect of splitting a scorecard

by sex and conclude that the Equal Credit Opportunities Act prevents improvements in scorecard performance. Other authors such as Crook, Thomas and Banasik (1995) explore the effects of splitting by other variables. They conclude that selection of a splitting variable must be carried out carefully to ensure improved results. In Chapter 4 we discuss the use of splitting scorecards and propose an approach based on a more rational argument which necessarily should lead to larger differences in predictive power.

1.1.5.6 Discretising Continuous Variables

Continuous variables are often categorised into discrete variables. Inevitably this process results in loss of information. Coarsely grouping ranges of continuous variables into discrete levels may conceal population changes that could otherwise have been taken advantage of by the scorecard.

1.1.5.7 Reject Inference

In credit scoring, scorecards are often constructed using only data from the accepted population, that is, those applications which were granted their requests. The resulting scorecard developed on these data is used to assess whether the population of new applications should be granted credit. These data have already been subjected to one selection mechanism. Since the scorecard constructed is to be used on the whole future population that applies for credit, modelling using only data from accepted applications results in estimation bias.

Rejected applicants never received credit and so their true class is unknown. Reject inference techniques attempt to assign probabilities to the rejected applications such that this information may be combined with that from the accepted population. The aim is to derive a scorecard from the data with the inferred information that is less biased than if no information on the rejects had been included. One method in frequent use is to include information from accounts that were rejected by the previous scorecard. The subsequent analysis

assumes that these rejected accounts would, had they been granted loans, have turned out to be undesirable accounts.

Reject inference is an important issue in credit scoring. However, throughout this thesis, we do not concern ourselves with this extra complication. Our aim is to identify and make improvements on other problems inherent in credit scoring. Hand and Henley (1993) give a critical discussion of reject inference methods. Joanes (1993/4) and Hand and Henley (1994) provide examples and discussion of reject inference in logistic regression and discriminant analysis.

1.1.6 Scorecard Construction

In this section we provide an overview of how a scorecard is constructed in an commercial environment.

The scorecard is based upon a classification rule that is estimated using the design set. The test set is used for assessing the likely future performance or detecting problems of *overfitting*.

Overfitting occurs when a classification rule constructed for future predictions discriminates with a high degree of accuracy between classes in the design set, yet the performance deteriorates substantially when assessed on the independent test set. Optimistically biased results on the design sample are produced when overfitting occurs. This problem is reduced when dealing with large data sets and so is unlikely to represent much of a problem in the credit scoring context where many thousands of observations are frequently used.

It is common practice for all variables, including continuous, to be divided into several levels. Many of the variables are already in categorical form, indicator variables arise naturally from application questions such as 'Do you hold a cheque guarantee card?' or 'Is credit protection required?' Other categorical variables arise from questions connected with loan purpose, bank accounts held or occupation. If numerous levels are defined they are often further grouped to

reduce the number of levels included in the analysis. Levels are grouped whose potential for default is similar.

A subset of the available variables is chosen to make up the predictors used in the construction of the scorecard. These variables may be chosen, subject to legal requirements, by some variable selection criterion, or by using prior knowledge from a domain expert.

The outcome variable, often some level of risk associated with the application, is then modelled by the selected predictor variables using a statistical technique. Typically in the industry this technique will be linear or logistic regression, although all manner of classification techniques have been applied to credit data sets. See Hand and Henley (1997a, 1997b), and Thomas (1998) for details.

The resulting model is applied to the test set. The model can be used as a scorecard providing no problems of overfitting are apparent. The parameter estimates of the model are transformed to provide rounded numbers which aid scorecard interpretation. These rounded numbers are the scores that are assigned to different variable levels and summed to produce an application's final credit score. To obtain scorecards of this form, additive linear models are required. Many non linear methods such as neural networks cannot be written in this form, consequently credit practitioners do not favour these approaches.

A manual check of the scorecard is then crucial. If one particular variable had associated scores far greater than the others, then the whole exercise of assessing new applications is reduced to an applicant's category for one variable. Additionally, the spread of attainable scores assigned to different levels of a variable should be sufficiently large to warrant inclusion of that variable in the scorecard.

Credit companies will constantly monitor scorecard performance. If the performance is not adequate then a different model will be estimated. This new model may have variables added or removed, or the way in which categories of

some variables are grouped may be altered. If the new model produces a superior scorecard then it may be adopted for use in the credit marketplace.

Figure 1.1 gives a graphical representation of the stages involved in scorecard construction.

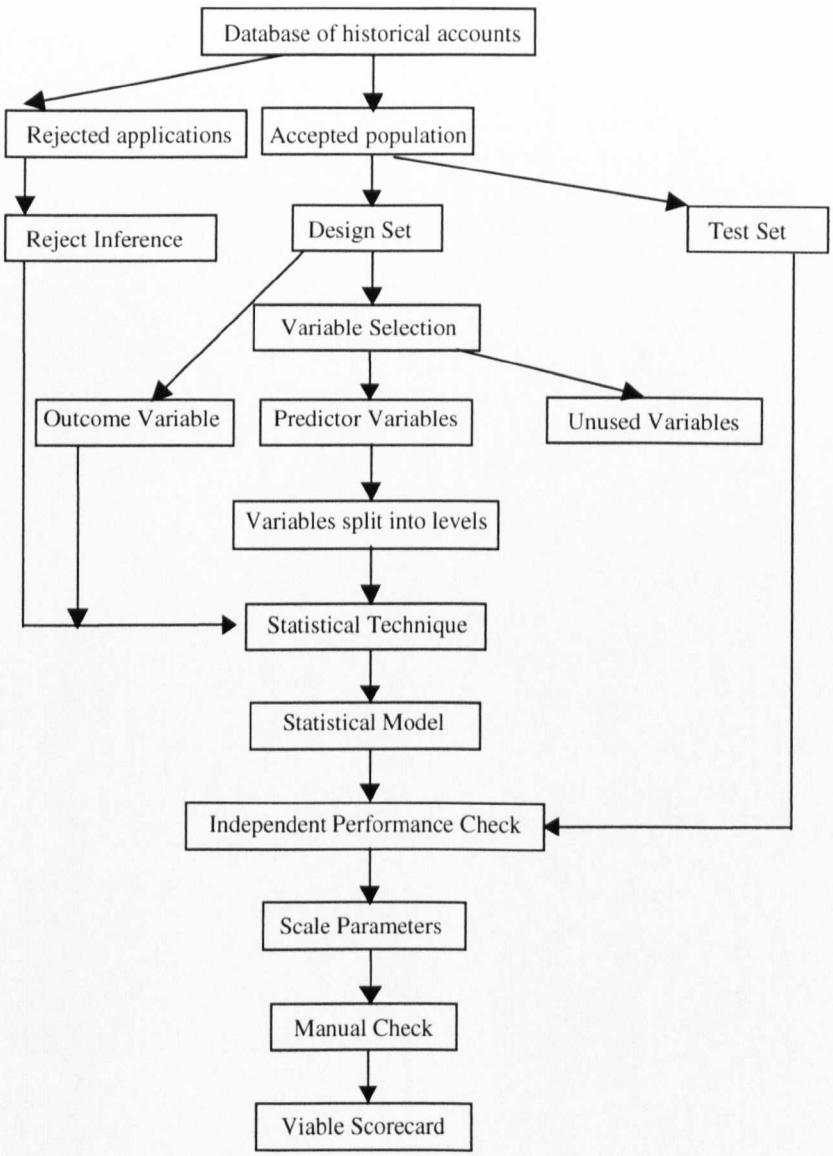

Figure 1.1: *A flow diagram of scorecard construction.*

1.2 The Credit Literature

In this section we will discuss the manner in which research in the area of credit scoring has evolved. We will concentrate on discussing the newer areas in credit scoring which have seen much recent research effort.

Hand and Henley (1997a) give a comprehensive review of the credit literature. Eisenbeis (1978) describes how the credit scoring literature is split into two parts. The first is concerned with business objectives. The primary concern of the second part is the application of different classification techniques and the relative benefits that each may exhibit in the resulting scorecards.

Many publications in the credit literature are concerned with achieving improvements in scorecard performance. These improvements can arise from different strategies. Using a different type of statistical classification tool, changing the variable selection strategy, altering the way in which categorical variables are grouped prior to analysis, are some of the possible ways in which performance can be enhanced. However, these improvements are often small. In this thesis we contend that other sources of variability, which are frequently ignored, often swamp the small improvements.

According to Henley (1995) linear discriminant analysis and logistic regression are the most commonly used techniques in the industry. However, credit analysts are continually investigating new ways in which predictions can be made with improved accuracy. Credit is an industry where improvements, even when small, can represent vast profits if such improvements can be sustained.

Recently, topical techniques such as neural networks and genetic algorithms have received much attention. For example Desai, Crook and Overstreet (1995), Arminger, Enache and Bonne (1997), Khoylou and Stirling (1993) discuss neural networks applied to credit scoring and Albright (1997), Fogarty, Ireson and Battles (1992) provide applications of genetic algorithms.

Other techniques, such as survival analysis have seen very little research activity. Narain (1992) highlighted its obvious potential by briefly demonstrating that a temporal element could easily be introduced into the credit scoring problem. The defaulters who are written-off very quickly are more costly than those who default towards the end of their credit agreements, yet time is rarely introduced into the problem. Banasik, Crook and Thomas (1998) conducted a more detailed investigation comparable to a real banking situation. Survival analysis will be discussed in further detail in Chapter 4.

It was recognised very early that Markov chain methods are a natural choice for some credit applications. Cyert and Thompson (1968), Cyert, Davidson and Thompson (1962) and Liebman (1972) produced early work in this area. Frydman, Kallberg and Kao (1985) give a more recent perspective.

Other credit problems have been investigated using a wide variety of statistical methods.

- Cluster analysis, Lundy (1992) Edelman (1992).
- Mathematical programming, Glen (1997), Gehrlein and Wagner (1997).
- Recursive partitioning methods, Marais, Patell and Wolfson (1984), Frydman, Altman and Kao (1985).
- Graphical models have also been studied. Hand, McConway and Stanghellini (1997), Sewart and Whittaker (1998), Stanghellini, McConway and Hand (1999).

Graphical models can be used for prediction. They may also be used to gain insight into the relationships between variables. If understanding of credit data sets can be improved, sources of uncertainty in the credit problem may be eliminated.

The sheer size of the data sets involved in credit scoring has recently resulted in applications of credit scoring appearing in the data mining literature, see

Feelders, Loux and Zand (1995) and Kelly, Hand and Adams (1998), Nakhaeizadeh, Taylor and Lanquillon (1998) for examples.

There are numerous publications devoted to assessing the relative methods of different classification techniques on credit data. Not surprisingly, these studies fail to generate a coherent message as to which method should be adopted as the gold standard for credit scoring. Each financial institution has different data, class definitions and credit products, so it is reasonable to presume that a different classification technique may be preferred on different problems.

Yobas, Crook and Ross (1997) compare neural networks, genetic algorithms, decision trees and linear discriminant analysis for a credit data set. They conclude, 'We found that the predictive performance of linear discriminant analysis was superior to that of the other three techniques. This is consistent with some studies but inconsistent with others.' Throughout the literature such findings are commonplace: no single classification technique consistently outperforms any alternative proposed. Indeed, Friedman (1995) states that no classification method is universally better than any other. This point of view is reinforced by the StatLog project (conducted between 1990 and 1993), with the aim of comparing many different classification techniques, on many different real problems including several credit data sets. Michie, Spiegelhalter and Taylor (1994) provide a text derived from the StatLog project.

However, with credit scoring we see a slightly different phenomenon. It is not simply that one classification method is not superior, but all classification methods perform very similarly. Considering different classification techniques applied to credit data, Thomas (1998) states that it is 'surprising that the recurring theme is how little difference there is in the classification accuracy of the different approaches'.

Possible reasons for such inconsistent and indifferent results may be:

1. Problems in credit scoring can be very diverse. For example, consider two loan data sets, one containing application information on the population whose applications were declined in addition to those whose requests were granted. The second data set consisting only of data concerned with accepted applications. The populations present in the two data sets differ substantially so it is plausible that different classification techniques better suit each of these situations.

2. It is often the case that an author has particular expertise in a certain area and so the comparison performed is biased towards their preferred technique (Ripley, 1996). Hand and Henley (1997b) state: 'Classification techniques can be compared in two situations: (i) as used in standard "commercial" applications, by someone who is competent but not a leading researcher in the methodology being used, and (ii) as used by someone who is actively involved in developing the methodology and will take advantage of subtleties of which the commercial user may be unaware.' The authors conclude that it would not be surprising for different results to be obtained by the two approaches.

3. In our experience of many different credit data sets, the classes in credit data are inherently inseparable. Our results are comparable with other publications in the credit literature, and it therefore seems reasonable to assume that our data sets are typical of what may be found elsewhere in the industry. The poor structure in the data can be modelled with similar success regardless of the classification technique used.

4. The flat maximum effect described by Lovie and Lovie (1986) and Overstreet, Bradley and Kemp (1992) suggests that a wide range of parameter estimates obtained by a linear model will give very similar results. This may be the reason that model changes often do not manifest themselves as changes in classification performance.

Some of the work in these areas has led to improvements, yet these improvements are often small. It is our opinion that benefits in performance from many model refinements are often eroded by other sources of variation inherent in credit scoring. For these reasons much of the recent research effort in credit scoring has focused on different approaches to solving the credit problems, taking the problem back to basics and thinking of new ways in which to address the problem rather than trying to improve on the existing methods. Leonard (1998) states that 'The most pertinent question regarding credit scoring is: 'what am I trying to predict?' This is a question that has frequently been asked in the industry and has directed company efforts towards modelling profitability.

1.2.1 Profitability

Early work in credit scoring often concentrated on modelling risk of default. The working assumption was that individuals associated with high risk would be most likely to experience difficulties when repaying the loan and be most likely to result in loss making accounts. Initially, using risk as a proxy for profitability worked well. However, with increased competition in the credit marketplace the industry has begun to tackle the complex problem of modelling profitability directly; see for example Thomas (1992a) or Hopper and Lewis (1992a). Although many organisations have experimented with profitability, in our experience implementation is still uncommon. One problem when attempting to model profitability is obtaining the necessary data to determine whether a customer is profitable. Thomas (1998) states that companies offering a single product may have insufficient information on their systems to confirm whether an account is actually profitable.

Working directly with profitability can often produce situations which may seem counterintuitive. Several examples of this are detailed below:

1. Accounts that may have traditionally been regarded as bad may in fact be the accounts which produce substantial profits for the bank. Consider loan

accounts. Account holders who keep up to date with the agreed repayment schedule are certainly good customers and provide a low risk and modest profits. However, high profits may be gained from customers who default on some monthly repayments, hence incurring extra charges. Oliver (1992) states that rejecting low risk applicants in favour of high risk applicants may lead to larger profits.

2. One may have regarded a portfolio of accounts each with long standing relationships as desirable. However, the competition amongst companies vying for the available business for products such as mortgages or credit cards is fierce. It may be the case that long-standing accounts are in fact the undesirable cases – those customers who have not been able to obtain better deals elsewhere.

3. Credit card account holders who pay-off their balance each month are low risk. Traditionally this is a desirable property for an account and can still lead to profitable accounts, as in point 1 above. However, with credit cards the profit associated with these 'no risk' accounts is reduced to a trivial amount. Thomas (1992a) suggests that the profitability of such accounts may hinge on whether an annual fee is charged on the account. Profitable accounts are those customers who maintain a credit deficit which is subject to interest payments.

1.2.2 Scoring Systems for Other Purposes

Behavioural scoring has been developed more recently than application scoring. Behavioural scoring is the monitoring of an existing account base. The aim of behavioural scoring is to facilitate the partitioning of the account base into several levels according to their associated propensity for some event. Imminent default is of particular interest, but not exclusively so. Other aspects of an account may be monitored so that credit terms may be changed where appropriate, or particular segments of the account base can be targeted with marketing campaigns. Since its introduction, the use of behavioural scoring has become increasingly popular. Chandler and Coffman (1983) and Hopper and

Lewis (1992b) discuss applications of behavioural scoring. Although used less frequently than application scoring the potential of behavioural scoring is great. The financial sector has begun to realise this potential, (Scallan, 1997). Behavioural scoring is often applied to current account portfolios. Using this information, accounts may be tailored according to the individual:

- Whether to offer the customer extra services.
- Whether to pre-approve applications for particular products.
- At what limit to set a customer's allowable unauthorised excess. That is, the amount overdrawn above their predetermined limit.

Performing such tasks without the knowledge of customers can produce great benefits. Primarily, business can be substantially increased. If products are pre-approved then many customers will choose to accept the offer rather than consider taking their business elsewhere. In addition customers are satisfied with the seemingly individual service. For example, a customer requesting an overdraft is likely to have some motivation for making the request. If behavioural scoring has predetermined that this account is suitable for such a facility and the account holder can be given an instant decision it is likely that the customer will be happy with the existing service and less likely to defect to a rival credit provider.

However the use of behavioural scoring is still in its infancy. Often information is collected on the conduct of accounts to enable a subgroup to be 'flagged' as bad, yet no action is taken against such accounts.

Credit scoring techniques are also used in fraud scoring. Essentially this is applying the same statistical classification techniques to a different problem. Leonard (1993a, 1993b) describes a scoring system designed to combat fraud in credit cards. Henley (1995) gave a brief comparison of scorecards specifically constructed for an application scorecard and one specifically built to identify fraudulent applications. Henley concluded that results, on his data set, were

similar whether fraud or risk definitions were used. Langley (1997) provides an account of a successful implementation of fraud scoring in a mail order company.

1.2.3 Innovative Models and Techniques

The main thrust of this thesis is to identify new areas in which improvements in the overall accuracy of a consumer credit classification system can be made and maintained. Ideas for alternative modelling strategies have long since been recognised yet few have been developed further than the concept stage.

What does it mean to say that someone is creditworthy? A definition must be proposed such that accounts can be classified as good or bad. These definitions will vary amongst banks and also of course between products. Crook, Hamilton and Thomas (1992) recognised that different definitions of creditworthiness may be used and compared the results obtained when using two different definitions. The results show that the choice of definition can have significant effect on the discriminatory ability of the resulting scorecards. Kelly, Hand and Adams (1998) take this idea further. They define a region of plausible definitions and investigate the performance of a logistic discrimination rule built for each of many definitions. This work is described in detail in Chapter 6. Both these papers suggest there is potential for improvements in classifier performance of magnitude far greater than improvements gained from fine tuning existing scorecards. Kelly and Hand (1997) proposed a class of models which enables a change in the definition used without the need for a complete recalculation of the scorecard, as described in Chapter 5.

The lifetime of a scorecard is also of interest. A scoring system is always based on historical credit data. Ideally we would hope that the population of accounts used to construct the classification rule and the population of applications for which future samples are drawn would be identical. However, with a time lag between the current applications and the historical data it is inevitable that some changes will occur. Consequently, once a scoring system is installed, its performance is likely to deteriorate until such a time that a replacement is deemed

necessary. Bazley (1992) advocates continuous updating of scorecards. Nearest neighbour methods have been explored for this purpose, see for example Hand (1997) or Taylor, Nakhaeizadeh and Kunisch (1997). A dynamic model, which updates as the class of each new observation arrives, is proposed and discussed in Chapter 7.

For further reading concerned with recent developments in credit scoring research see Hand (1998) and Thomas, Crook and Edelman (1992).

1.3 Scorecard Assessment Criteria

In this section we will discuss several methods used for assessing the performance of classification rules. Wilkie (1992) describes other measures that may be used for comparing scoring systems. Credit problems addressed throughout this thesis will be treated as two class classification problems. This is the most common approach in the industry. The general confusion matrix for a two class problem is presented in Table 1.1.

For the two class problem, we denote the good risk class by class 0 and the bad risk class by class 1. Many classification methods output $\hat{p}(1|\mathbf{x})$, an estimate of the probability of belonging to class 1 at $\mathbf{x}$, where $\mathbf{x}$ is a feature vector. This estimate is compared with a threshold in order to form a classification rule. This rule assigns observations to class 1 if $\hat{p}(1|\mathbf{x}) > s$, and otherwise to class 0, where s is the classification threshold.

		True Class	
		0	1
Predicted Class	0	*a*	*B*
	1	*c*	*D*

Table 1.1: *Two Class Confusion Matrix*

We denote π_0 and π_1 as the observed prior probabilities. π_0 is the probability of a 'good' customer and π_1 the probability of a 'bad' customer. p_0 and p_1 are defined as the analogous predicted probabilities.

Below we discuss three different assessment criteria. First, error or misclassification rate, that is typically used in areas where research on classification methodology is ongoing. The second criterion is based on the receiver operating characteristic (ROC) curve. This method is commonly used throughout the part of the classification literature concerned with practical applications. The third criterion is loss. Although perhaps the most appropriate criterion for comparing or assessing classification rules it is seldom found in the literature.

1.3.1 Error Rate

Much research effort has gone into error rate estimation. Hand (1986) discusses several types of error rate. The *Bayes error rate* (Fukunaga, 1990) is the best achievable error rate possible for a given set of variables. In practice this is extremely difficult to estimate. The *true error rate* is defined as:

$$e_T = \int_{\Omega_a} p(1)p(\mathbf{x}|1) + \int_{\Omega_r} p(0)p(\mathbf{x}|0)$$

Where Ω_a and Ω_r are the accept and reject regions, $p(0)$ and $p(1)$ are the good and bad design set priors and $p(\mathbf{x}|0)$ and $p(\mathbf{x}|1)$ are the class conditional probabilities.

When an independent test set is available an unbiased estimate of the true error rate may be obtained from Table 1.1. This independent estimate of error rate is defined as $\frac{b+c}{a+b+c+d}$, that is, the number of incorrectly allocated class 0 and class 1 objects expressed as a proportion of the total number of observations. In credit scoring sufficient data is usually available to enable the use of an

independent test set. However, when this is not the case care must be taken to ensure unbiased estimators are produced. Toussaint (1974) and Hand (1986) detail suitable methods. Efron (1983) discusses a bootstrap approach to estimation.

Error rate assumes that misallocation costs are equal, that is, the loss incurred when incorrectly classifying a bad is the same as misclassifiying a good. However, this is rarely the case in practical classification applications. For any classification problem different versions of Table 1.1 can be obtained by choosing different values of classification threshold. An ROC curve is a graphical representation that encapsulates the information obtained by varying the classification threshold across all possible values of threshold. ROC curves are discussed in the next section.

Misclassification rate is often used in the credit industry. However, it is not sufficient to minimise misclassification rate. Consider a credit application where only 5% of the population form the smaller class. A classification rule that assigned all observations to the large class would give a very favourable misclassification rate. However, the credit grantor may prefer to reduce the number of accepted bad accounts at the cost of reducing the number of goods accepted. This gives rise to an approach often used in the credit industry. That is, specifying the proportion of applications that will be accepted and minimising the bad rate amongst those accepted. Note that one consequence of this approach is that bounds on the performance are imposed. For example, the best achievable bad rate among the accepted population is given by:

$$Min_{BR} = \begin{cases} 1-(p/a) \text{ if } a > p \\ 0 \qquad\quad \text{if } a < p \end{cases}$$

Where a is the predetermined proportion of the population accepted and p is the proportion of true bads in the population. Henley (1995) discusses this performance criterion.

1.3.2 ROC Curve Analysis

Receiver operating characteristic (ROC) curves are commonly used in financial and medical applications. A common variant of an ROC curve is obtained by plotting *true positive rate* $(\text{TP}) = a/(a+c)$,on the vertical axis against 1-true negative rate, where *true negative rate* $(\text{TN}) = d/(b+d)$, when the confusion matrix is evaluated across the full range of thresholds. True positive rate is equivalent to the proportion of goods accepted and true negative rate, the proportion of bads rejected. For the most part we will plot ROC curves with axes 'bads accepted' against 'goods accepted'. ROC curves are sometimes referred to as Lorentz diagrams.

Considering the cumulative distributions of predicted good and bad accounts, this plot is defined as the cumulative proportions of good and bad accounts with predicted probability less than or equal to some classification threshold *s*, plotted against one another. A typical ROC curve is illustrated in Figure 1.2.

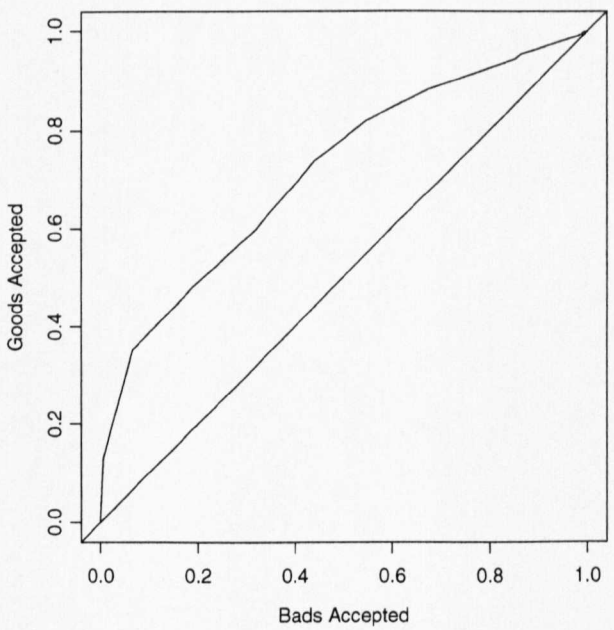

Figure 1.2: *An ROC Curve*

The curve shows the relationship between the number of goods rejected and the number of bads rejected. A suitable classifier can be chosen by selecting an appropriate point on the ROC curve. The closer the ROC curve is to the top left corner of the square the better the classifier at discriminating between objects. For a given cost the optimal point on the curve can be found by a simple projection as described by Hand (1997, p133).

Several measures may be derived from the ROC curve. The most commonly used in the financial area is the Gini coefficient, (Hand, 1997). The Gini coefficient is defined as twice the area between the ROC curve and the diagonal. The Gini coefficient takes values on [0,1] classifiers with larger Gini coefficient are considered superior. However, the Gini coefficient summarises classifier performance over the whole range of classification thresholds and so should not be regarded as an absolute measure of performance. Problems with the Gini coefficient and error rate have been highlighted by Adams and Hand (1998a).

The area under the curve (AUC) is a simple transformation of the Gini coefficient. $AUC = \frac{G}{2} + 0.5$, where G is the Gini coefficient. As its name suggests the AUC is the area of the ROC square underneath the curve. Similarly the larger the AUC the better the classifier is at discriminating between classes. Both the AUC and Gini coefficient are easily calculated using simple numerical integration such as the trapezium rule.

1.3.3 Loss

The cost of classifying a class 0 object as a class 1 object, c_0, (i.e. a good credit account as a bad credit account) is demonstrated in the form of revenue lost that would only have been present if the account had been accepted. The opposite misclassification, c_1, that of classifying a bad account as a good is more serious. In this case the cost occurs as an actual financial loss. The credit account is not

repaid in full accordance with the credit terms agreed. When these costs are known, a measure of loss is given by

$$L = \pi_0 c_0 (1 - TP) + \pi_1 c_1 (1 - TN)$$

In financial applications loss would certainly seem the most appropriate measure. Loss could enable classification performance to be quantified in terms of profit. If one classifier makes more profit than a competitor then certainly in some sense it would be regarded as superior. However, difficulties arise in implementation. The costs associated with misclassifying accounts are often difficult to obtain reliably.

Loss in general is more complicated. The approach presented above assumes that correct allocations have zero costs associated. However, this is certainly not true in an application of credit scoring. Correctly classified good accounts may incur a processing overhead. Similarly, correctly classified bad accounts may have warranted expensive credit checks.

1.3.4 Recent Improvements

None of the measures presented above are ideal in all situations. Indeed, different problems are always likely to warrant different approaches. Recently several improvements have been proposed. Moreover, the classification community has begun to realise that classification performance is not easily summarised by a single number such as error rate or Gini coefficient. Instead, detailed knowledge of the problem and data are crucial in order that sensible conclusions be drawn. Salzberg (1997) provides insight into common problems that occur in large sample analysis and details some common pitfalls that can easily be overlooked yet render results invalid.

Much research activity is focused on classification rule comparison. In this section we briefly discuss some recent progress in the area. However, the most common method of assessment used in financial institutions is ROC curve

analysis. For this reason the remainder of this thesis presents results using the traditional methodology, misclassification rate, ROC curve analysis and loss.

1.3.4.1 Loss Comparison Analysis

Adams and Hand (1998b) propose a new criterion for comparing different classifiers, the Loss Comparison index. This idea is based on a plausible range of costs which can usually be determined even when precise costs are unknown. The loss is then calculated for each threshold value in the range of plausible costs. The result is a projection of an ROC curve which is integrated between the range of costs to determine which classifier is to be preferred. This approach may be preferable to ROC curve analysis when comparing classifiers because some knowledge of costs is incorporated and attention is focused only between the likely range of costs rather than summarising across the whole range.

1.3.4.2 Modified ROC Curve Analysis

ROC curve analysis does not account for class priors. Often this causes no problems as the class priors are fixed. However, when comparing many different models based upon different class definitions, as is done later in this thesis, the class priors can alter substantially. In order to account for the priors we must modify the ROC curve.

Geometrically we may regard the top left corner of the ROC square as the origin. The vectors $(1\text{-}TN, 1\text{-}TP)$ and (qc_1, pc_0) define directions from this origin, where $p=1\text{-}q$ is the probability of belonging to class 1. The vector (qc_1, pc_0) remains unchanged for different rules and so the loss is determined by the location of $(1\text{-}TN, 1\text{-}TP)$ on the ROC curve. For two rules **a** and **b**, the inner product of the two vectors (qc_1, pc_0) and $(1\text{-}TN, 1\text{-}TP)$ defines the overall loss. For two classification rules, rule **a** is superior if this inner product is less than that of rule **b**. This is demonstrated geometrically by the ROC curve for the superior rule being closer to the 'curve' joining (0,0), (0,1) and (1,1) for each point on the axes.

If the class priors change, then the vectors (qc_1,pc_0) also change, this complicates the problem. The ROC curve can be generalised by plotting $q(1\text{-}TN)$ and $p(1\text{-}TP)$ as axes instead of $1\text{-}TN$ and $1\text{-}TP$. The ROC curve will now be confined between the limits $[0,1\text{-}TN]$, on the x axis and $[0,1\text{-}TP]$ on the y axis. Interpretation is the same as standard ROC curve analysis, although based on the inner product of $(q(1\text{-}TN),p(1\text{-}TP))$ and (c_0,c_1).

Two standard ROC curves may have been virtually indistinguishable plotted in the usual way. However, a distinct advantage of these modified plots is that the curves from two rules must cross, and so one rule is seen to dominate for certain values of c_0 and c_1. For more details see Hand and Kelly (1998).

1.3.4.3 The ROC Convex Hull Method

When comparing classification rules using ROC curve analysis Provost and Fawcett (1997) propose using convex hulls to aid assessment. The data used to plot the ROC curves for the different classifiers is pooled and a convex hull computed for these data. For relevant costs, a projection can be made onto the ROC square. This results is a line where all classifiers corresponding to points on the line have the same expected cost. The point on the projected line that intersects the convex hull and has the largest proportion of goods accepted (TP) gives the optimal classifier.

1.3.4.4 Realisable Classifier Theorem

Given that a fixed proportion of bads will be accepted, we may chose the classification rule that achieves the largest proportion of goods. This is a common way of choosing a classification rule in credit applications. ROC curves are often not smooth when estimated from small samples. When this is the case the point on the ROC curve chosen by the above method may represent a conservative proportion of goods accepted. The realisable classifier theorem (Scott, Niranjan and Prager, 1998) constructs a randomised decision rule based on the information encapsulated in the ROC curves being compared. This theorem can lead to a

superior classification rule when comparing classification rules whose ROC curves cross.

Chapter 2

Data: Structure and Features

2.1 Introduction

Data are crucial at all stages of any credit scoring procedure. Handling and processing of data may influence the resulting model as much as the choice of variables included in the model. Irregularities in the collection phase may lead to incorrect conclusions being drawn. Data utilised in the construction phase that is not representative of the current population is likely to lead to an unreliable scoring system. Large databases are usually used when constructing scorecards.

Classifiers are constructed and compared for several data sets throughout this thesis. The two credit data sets used for this purpose are described in this chapter. These data sets are real banking examples. The first is a database of unsecured personal loans which will be used for application scoring purposes. The second is a portfolio of current accounts used for behavioural scoring.

2.2 Variables

Huge numbers of variables are often stored in financial databases, yet few used in scorecard development. High dimensional databases may create a variable selection problem which could be investigated as a topic itself. However, in this project we choose to concentrate on the comparison of techniques rather than enhancing a particular method by fine tuning the choice of variables included in

the model. Consequently we shall make use of a subset of the variables, similar to those frequently used in industry for the development of scorecards.

As discussed in Chapter 1, credit scoring can take the form of application scoring, where the creditworthiness of new applications is assessed, or behavioural scoring, in which the performance of existing accounts is monitored. Categories of data are summarised below:

- Application Data: Information derived from the application form.
- Bureau Data: Information supplied by Credit Reference agencies.
- Behavioural Data: Information collected as real accounts evolve.

2.2.1 Application Data

Many data are derived from loan application forms. These data can be partitioned into one of the following categories.

1. **Personal data** includes variables related to the applicant (such as occupation and salary), the applicant's family and other financial commitments.

2. **Established financial data** consists of information in connection with credit agreements already formed. This can include other types of accounts held, pensions, credit cards or whether the applicant has previously applied for similar credit products.

3. **Product information** details the amount of the desired loan, its purpose, how it will be repaid and the length of the repayment period.

2.2.2 Bureau Data

Bureau variables typically consist of information connected with credit applications elsewhere in the marketplace or legal data concerned with past payment irregularities on previous credit agreements. Examples of such variables

are age of, or number of, County Court Judgements, customer account information sharing data, which provides details of the conduct of previous credit agreements.

2.2.3 Behavioural Data

Behavioural data may consist of any information connected with the conduct of an account. This can be payment information, numbers, frequency or cost of transactions or account services used. Behavioural data is collected mainly for two reasons. Firstly, customers whose accounts are well conducted may be offered extra services, or be targeted by new marketing schemes. Secondly, action may be taken against customers whose potential for bad debt is high.

2.3 Unsecured Personal Loan Data Set

In this section information on unsecured personal loans (UPL) is described. We use two sets of UPL data which arise from different time periods. The following description is applicable to both data sets. These are treated distinctly throughout the thesis because different sampling schemes (discussed later) were used to extract the data from larger databases.

The first, UPL(1), comprises loan accounts which first became active in 1992 and were collated in 1995. A maximum of 35 months of information concerned with the behaviour of the account, once activated, is available on accounts in this data set. The second, UPL(2), consists of accounts which became active between the years 1993 and 1996. For all accounts in UPL(2) performance data is available from the first loan repayment until November 1997. Depending on the year from which the loan originated, each account has 59, 47, 35, 23 or 11 months of associated data.

These UPL data sets consist only of information from the customers who were accepted. Scorecards are often constructed using information gained from

applications that were previously scored and accepted by the selection mechanism in place at that time. This practice leads to problems of bias (reject inference) as described in Section 1.1.5.7.

2.3.1 UPL Data Specifics

More than one hundred variables are stored in the UPL database, many of which are not useful when statistically analysed, such as applicant's initials or application numbers. However, the majority have potential for at least some predictive power. These variables seem to form two distinct subgroups, those whose contribution to a classifier may improve the predictive power of the model substantially (e.g. time in current job, county court judgements) or those which may have a slight impact on the overall performance of the model (e.g. branch of application, joint application).

Available variables that are classed as behavioural data are the *state* of the account, an indication of the status of an account, and *outbal*, the outstanding balance. These variables are recorded for each month that the loan account is active. A typical account from the UPL(1) data set could have up to thirty five month variables, *month1* to *month35*, each with a performance indicator, *state* defined below, and *outbal1-outbal35*, corresponding to the respective outstanding balances of the loan. The *state* that each *month* variable may take is determined by the payment history of that account. Table 2.1 lists the meaning of the levels of *state*.

Meaning	STATE
All payments made in accordance with the loan agreement	0
Between one and five monthly payments in arrears	1- 5
Six or more monthly payments in arrears	6
Legal proceedings for debt recovery initiated	98
Account written-off	99

Table 2.1: *Description of the behavioural variable* state.

The monthly replicates of *state* variables are used to derive a further variable, *worst*, that indicates the worst state of arrears attained by an account throughout the observed period. This worst variable is often utilised when deciding upon the definition of good and bad that is to be modelled.

Bureau data can be very predictive, often representing past irregularities on previous credit agreements. In this data set only customers who were accepted for a loan are included. These data have effectively been scored already by the scorecard that was used in the marketplace at the time of application. For this reason although bureau data is available it is not particularly beneficial to the discriminatory ability of models constructed from the available data.

2.3.2 Sub Sampling

The vast majority of account holders pay their monthly premium on time each month throughout the lifetime of their loan. Consequently, the percentage of accounts falling into arrears is relatively small. Due to this massive imbalance in class sizes some of the data available on the customers who never fell more than one month into arrears were excluded from the analysis. The resulting data set requires reduced computing time for analysis and the sub-sampling could be accounted for in the results by re-weighting the population priors. Tarassenko (1998, Appendix B) details how to update predictions when data has arisen from sources with different prior probabilities. Further research in this area is necessary to establish the best course of action when dealing with populations whose class sizes are vastly skewed. By means of simulation, Hand and Kelly (1998) show that Gini coefficient increases as the size of one class tends to zero.

2.3.2.1 UPL Data 1992 - Sampling Scheme 1

The first set comprises 23189 observations that were sampled from a larger population of 53799 applications who were successful when applying for unsecured personal loans.

One in three of the customers who never fell into arrears were randomly assigned to our data set. Similarly one in two of the customers whose worst arrears was one month were included. Accounts with arrears greater than two months were included in the data set. The resulting data set of 23189 observations is summarised, in terms of worst arrears (defined in Table 2.1) attained during the observed period. (Table 2.2).

Worst	Frequency	Percent	Sampling Fraction
0	13241	57.1	1/3
1	4128	17.8	1/2
2	1651	7.1	1
3	807	3.5	1
4	488	2.1	1
5	313	1.3	1
6 +	1082	4.7	1
98	403	1.7	1
99	1076	4.6	1

Table 2.2: *Summary of arrears for accounts in UPL(1)*

2.3.2.2 UPL Data 1993-1997 - Sampling Scheme 2

The second unsecured personal loan data set consisted of several hundred thousand records in total. In this case the selection scheme only retained 24 month loan accounts. This approach would remove some variation introduced to the problem when different loan terms are grouped (discussed in Section 4.3.1). All observations were included in our data sets for analysis if they ever fell into arrears by any amount. Forty percent of accounts with no arrears were also included in the data sets. The resulting data set consists of 92258 accounts.

2.3.3 Changes Made to the Data

There were many inconsistencies contained in the data set. The sponsoring bank could not explain all these occurrences and could only attribute them to errors in the computer systems monitoring the performance of active accounts. This section details such problems and indicates any corrective action taken.

Customers whose worst state of arrears was 'Legal proceedings' (*worst*=98) were sparse throughout the data. Legal proceeding is a very undesirable stage for an account to reach, so these accounts were combined with the write-off (*worst*=99) accounts and re-coded accordingly. When a customer's account is written-off the corresponding outstanding balance is set to zero because recovery of the outstanding debt is not expected. If the status of an account is still at the legal proceedings stage, the debt is actively being sought. Corresponding outstanding balances were set to zero for the duration that legal proceedings were ongoing.

Accounts that jumped from zero to write-off (*worst*=99), or from zero to legal proceedings (*worst*=98), were identified and removed from the data set. The common problem with these accounts was the final payment. A small amount of the true outstanding balance was unpaid, usually not much more than a few pounds. A likely explanation would be that the final payment was slightly greater than previous instalments but the account holders did not adjust for this difference. Such amounts are treated as negligible and disregarded. When no attempt to recoup the remaining outstanding balances was made the system coded such accounts as write-off. Only a small fraction of accounts followed this pattern. For example, the UPL(1) database contained 203 such accounts. When dealing with a large data set 203 accounts may be considered negligible, however, a total of only 1276 of the 23189 observations are classified as write-off. When considering the occurrence of write-off accounts, 203 observations represent a substantial proportion of the total. Dubious conclusions may result if this source of technical write-off were grouped together with the true cases of write-off. This point along with others which may influence results substantially, will be discussed further in Section 2.5.

Negative values in the variables holding outstanding balance information were commonplace. This was probably due either to an account holder overpaying their final instalment, paying extra instalments when the agreement was complete, or continued payment of the appropriate direct debit. These customers

would eventually have been refunded the amounts overpaid, so negative entries in the data set were set to zero.

In the UPL(1) data set twelve and twenty-four month loan accounts were deleted if the account was still active at the end of the thirty five month observation period. A plausible explanation is that the majority of these customers extended their loan terms. Alternatively these accounts may have been customers who were experiencing extreme difficulties yet were actively trying to repay. However, such information was not available so such accounts were removed. This class of observations was small, yet before deletion they were examined closely to ensure they were not of a particular type of customer.

Accounts whose outstanding balance became zero at month x and then became positive at some later month y, where $y>x$ were deleted as this kind of behaviour was unexpected and unexplainable, even by bank experts.

A small proportion of accounts in the database had all their outstanding balance entries as zero or missing. It was unclear whether these were simply errors in the data or whether they were another class of outcome - that of the customer who declines the offer of a loan. However, even if the latter supposition was true the numbers were not sufficiently high to enable reliable analysis to be performed. These observations were deleted.

In principle, monthly state is a variable that should only increase in single units at each month until such a time that the arrears are sufficiently high as to justify write-off. However, the monthly cycle adopted by the bank is a four week period which may run from the 1st to the 28th of the month. Monthly states in the data set however, are calculated on the basis of calendar months, so it is possible for some accounts' state to increase by two units at a single time point. These observations are retained. Observations whose monthly state variable increased by more than two units after only one month had elapsed were removed. Note that the same is not true of decreasing monthly state, once in arrears the bank will permit all arrears to be repaid at a single time point.

After consideration of the points presented in Section 2.3.3 The original data sets, UPL(1) and UPL(2) consisted of 22490 and 92258 accounts respectively.

The data modifications detailed above are legitimate, as our purpose is to explore and develop new methodology. However, if the ideas presented in this thesis were to be applied in practice it would be necessary to extend the models to incorporate such anomalous observations providing a more satisfactory explanation for the cause of such observations could not be ascertained.

2.3.4 Special Features of the Data

The unsecured personal loan data sets UPL(1) and UPL(2) have several peculiarities associated with them which may not be found in typical credit data sets. In this section such features will be detailed.

2.3.4.1 Criterion for Entry into the Data Set

The data collection procedure is the same for all available years of data. In this section we shall illustrate the collection phase by discussing the UPL data which originated from the 1992 cycle of loan application. The structure of these data makes it impossible to detect phenomena related to time, as illustrated in Figure 2.1.

Figure 2.1 shows examples of loan agreements that would be included in our data set. The solid lines indicate the time for which the loan was active. Accounts were included in UPL(1) data set if their loan agreements commenced in 1992. Regardless of the time of year at which the account became active the data connected with the first monthly payment would be entered into *month1* and so forth. The resulting effect is that any time structure embedded in the data will be lost. Also we have the same amount of monthly replicates for an account which began in January as for an account whose application was granted in December of the same year.

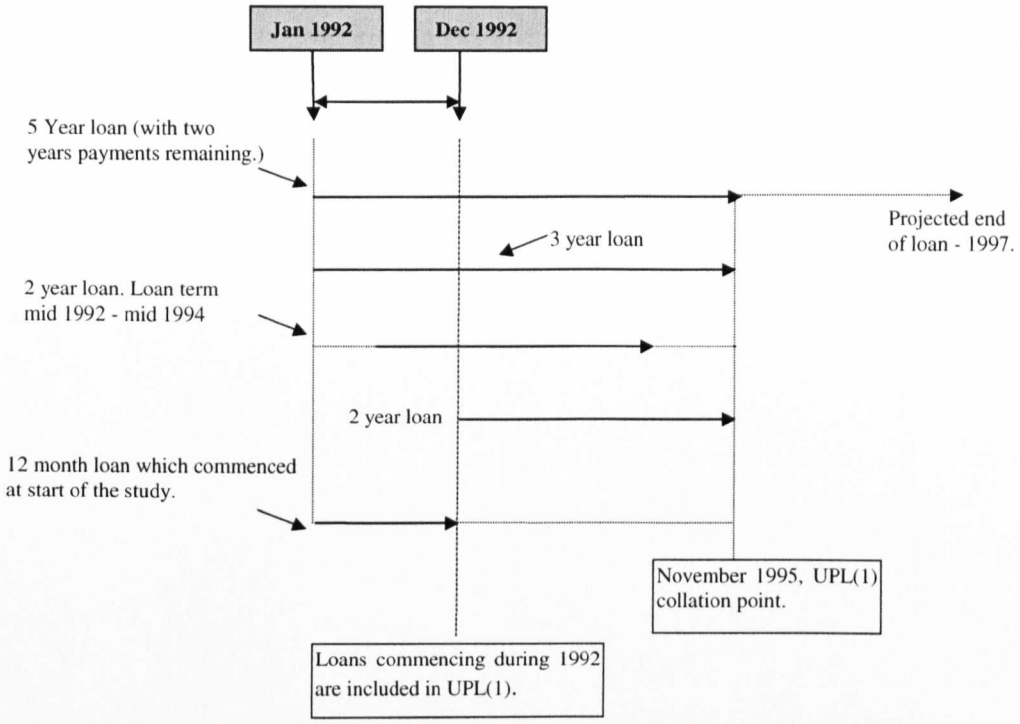

Figure 2.1: *The criterion used in deciding which accounts should form UPL(1)*

2.3.4.2 Ordered Trends

The data were recorded monthly and hence are ordered. Furthermore, the data are such that when a customer is falling further into arrears only certain outcomes can be attained at certain points, yet one can recover from several months in arrears in a single step. For example, if a customer has payments in arrears of say two months, then the only options available to them for the following month are:

1. Pay the following month as required and maintain the two months in arrears status.

2. Pay more than one month's premium and reduce the state of arrears.

3. Fail to pay the premium due and fall three months in arrears.

As already noted, these restrictions on state of arrears from month to month when a customer is falling into arrears do not apply when someone is reducing their arrears.

2.3.4.3 Truly Discrete Behavioural Data

In many medical situations the exact time at which an event occurs can often not be pinpointed. For example, patients may be required to attend a monthly check up in case of disease recurrence. If the disease does recur, the exact time of recurrence cannot be established precisely. One can only say that recurrence occurred during the month prior to detection.

This is not the case in the unsecured personal loan data. Although a similar framework exists, a customer with a loan cannot go a month further into arrears two weeks before the next payment is due. By definition missing a payment can only happen on one specific day of the month. This phenomenon gives rise to a set of truly discrete data.

2.3.4.4 The Issue of Censored Data

An observation is censored if the end point of interest is not observed (Cox and Oakes, 1984). In a credit context the end point being modelled may be the time when a credit account is written-off due to serious arrears. Censoring occurs if an account does not default during the observed data period but still has repayments outstanding.

Due to the way in which these data were collated, loans of period strictly less than the duration of available data should not be censored. However, in reality, due to missed payments and defaulters, a substantial amount of this data is censored. This is because the number of breaches of the agreed payment schedule extends the time that the account is active, hence the total duration is longer than the observation period. Active accounts with loan terms greater than their respective observation periods are censored.

In addition, customers whose accounts have life spans less than the observed period of data have two options. Customers can either repay their loan in full with few problems or may default and be unequivocally classed as 'bad' customers. Although highly undesirable, accounts conducted in this way cause no problems in the statistical analysis. Customers who pay-off their loans in accordance with their loan agreement are not censored in the way described above. Their agreement is complete and so they have no further possibility of defaulting. This is in direct contradiction with the assumption usually made about censored data, that is, the time of failure is unknown yet certain if the observation period is long enough. This different type of end point should be incorporated into the analysis.

2.4 Current Account Data

The data set comprises 27836 accounts and more than 500 variables. These data are predominantly a behavioural data set. Consequently, many variables arise from variables being replicated on a regular basis.

2.4.1 Current Account Data Specifics

Many application and bureau variables were included in the data set. However, this information will not be used because the accounts in this data set are those which have already been credit scored and granted a current account. Instead we focus our attention on a subset of 35 variables which vary with time. Monthly replicates of these data are stored. Examples of such variables are monthly end balance, credit turnover, number of direct debits and value of cheques presented to the bank during the previous month.

We predict a performance measure using these data. The performance measure used in the bank is complex with good, bad and indeterminate classes. Each class has up to six properties defined, any one of which determines class membership of an account if the account demonstrates that property. We implement a slightly modified version of the performance measure to preserve confidentiality and to

make the problem manageable in the time available, details of which can be found in Chapter 6. The performance measure is predicted n months in the future where n can be any number between one and twelve.

This data set had undergone some sampling procedure before use in the bank. However, such information was not available and so not incorporated into any analysis.

2.4.2 Changes Made to the Data

In order to use the data several modifications were required. First, accounts in specific groups defined by the bank but not involved in our performance definition were excluded. These included write-off accounts, accounts which had been referred to the collections department, those for which information was available but the account had been closed, new accounts which were less than one month old and accounts which had remained inactive for three months. The resulting data set comprises 16681 accounts. Secondly, we eliminated accounts with missing data. In the UPL data sets, missing values were included in the analysis to aid prediction. However, in the current account data set we are concerned only with variables monitored by the bank. These missing values can be attributed to (a) computer system shortcomings and (b) frequently accounts that were not open a sufficient length of time to give the required amount of data. The final data set consists of 12547 current accounts.

2.4.3 Special Features of the Data

The current account data sets were supplied pre-processed whereas the UPL data sets were in their raw data form. Consequently problems with these data had already been removed. The current accounts are not constrained to the same extent as the data concerned with repayments on a loan so particular patterns and issues of censoring do not arise. Given a longer data period it is likely that

interesting trends may be detected in variables with monthly patterns such as number of direct debits or cheques.

One interesting feature of the current account data was that for many variables the monthly means were consistently decreasing over time. This type of pattern could easily be overlooked at the analysis stage and yet have notable impact on results. When challenged on this matter the bank provided the following reasonable explanation. The decrease was due to a change in interest rate over a twelve month period (from 18% to 9.9%) coupled with a change in banking policy which effectively increased the proportion of authorised overdrafts in the current account population.

2.5 Problems Arising from Large Data Sets

Large data sets can often be problematical to analyse. Many factors can lead to difficulties. The computational power necessary may be an issue, data maybe distributed across several systems or data may be difficult to visualise. In this section we describe several instances of data irregularities which could be the cause of incorrect results. Many traditional techniques may be of limited use when dealing with many thousands of data points. In certain circumstances conclusions drawn may provide misleading results.

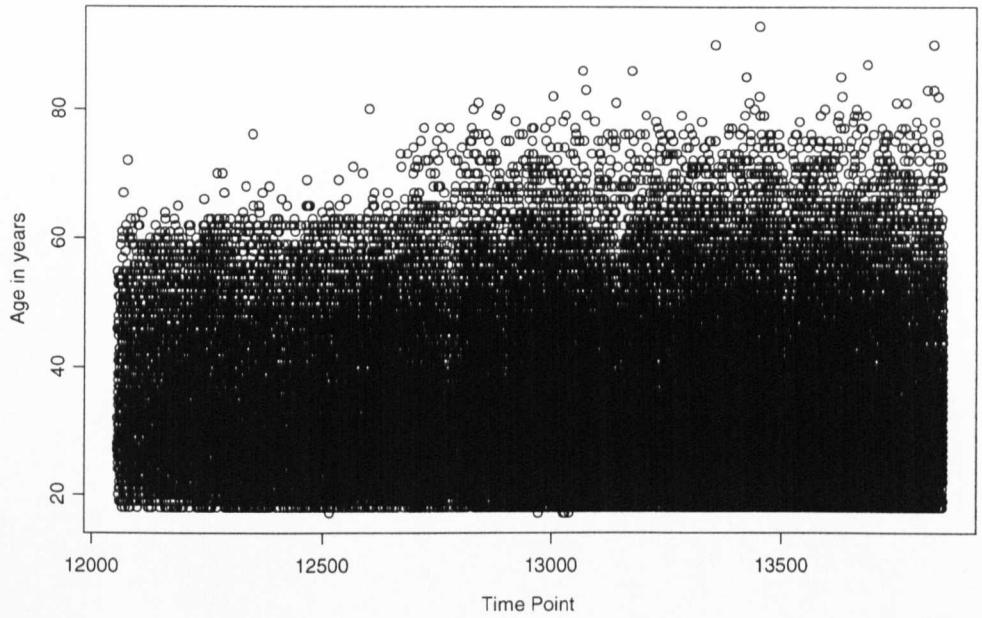

Figure 2.2: *Large Sample Scatter Plot*

Figure 2.2 shows a simple scatter plot of 92258 current accounts. This plot is very uninformative. A dense black area is the result of many overlaid points. Superficially, it may appear as if the underlying trend is increasing. In fact the opposite is true. When a regression line is fit to these data we see a significant result in the form of a decreasing trend.

Misleading results can be obtained when dealing with large data sets simply because the amount of data is so vast that large discrepancies in the data go unnoticed. Note that when dealing with large data sets, statistically significant results are more likely. The chance of overlooking data irregularities is at its largest when automated analysis is performed. The errors discussed in this section could easily, and perhaps have, passed undetected, through the process of constructing a scorecard in a financial institution. In an industry where small changes in classifier performance can equate to huge profits these examples highlight the consequences of working with a database which is neither precise nor consistent.

Note that the database errors reported in this section are from several data sets to which we have previously had access. These examples are collected below for illustrative purposes and do not necessarily relate to data sets used in this thesis.

2.5.1 Periods of Missing Variables

Credit scoring typically involves building a classifier from historical data, which is used to determine the suitability of future applications for credit. The scorecard built may then be used in the marketplace without modification for long periods. Common practises include grouping levels of a categorical variable into fewer levels such that a score may be associated with each. Frequently missing values may be regarded as informative and as such incorporated into one of the categories.

Upon close inspection we found that one of the most influential variables in the scorecard was consistently recorded as missing for a twelve month period. Concealed somewhere in the hierarchy of the financial institution there may lie a sensible explanation why this is the case. However, it is unlikely that attention would explicitly be drawn to this when the time came for development of a new scorecard. The period where the variable was missing would most certainly be detected if it occurred in the period of data used in scorecard construction. However, should the period with the missing variable correspond to a time shortly after a new scorecard had been implemented, it may be undetected for a considerable time.

A wide range of scores may be associated with the different levels of a variable. Implementation of a scorecard in the credit marketplace when the period of accounts with missing variables occurred could prove extremely problematical. The resulting effect would be to lower all application scores by the maximum spread of the obtainable scores of the missing variable. The consequences of such an error may be that many thousands of pounds worth of business are rejected

which ordinarily would have been accepted due to applications falling just below the cut-off score in place at that time.

2.5.2 Incorrectly Coded Observations

A customer is often defined as a bad credit risk if their account falls a fixed number of months in arrears or if they fail to repay the outstanding debt by an agreed time. Definitions such as these coupled with automated systems often lead to irregularities when logging data on a large customer database.

In Section 2.3.3 it was noted that a typical unsecured personal loan would involve a fixed monthly repayment for an agreed number of months followed by the final payment which settles the balance. Further investigation revealed that some accounts which had just a few pounds outstanding were being coding as bad regardless of how the account was conducted.

Typically credit problems deal with small and difficult to predict bad classes, so even a small proportion of these 'technical bads' could adversely effect results.

2.5.3 Potentially Influential Data Patterns

One continuous part of a data set had about 10% of the number of bads that would normally be found a data set of this type and quantity.

If the data were taken collectively and used for scorecard construction then, the fact that the first portion of data was unusual may go unnoticed. Consequently, the scorecard constructed from this data would be estimated using a data set that was not representative of the true populations. In such circumstances a sub-standard scorecard would not be surprising.

Chapter 3

Current practice

3.1 Introduction

Before a decision to offer an application credit can be made, there must be some notion of what is meant by creditworthy. A set of rules must be determined to construct a definition of creditworthiness.

When assessing applications for unsecured personal loans the population of applicants is partitioned into distinct subsets using definitions similar to the following:

A customer is 'bad' if they have ever been three months in arrears, or are currently two months in arrears.

A customer is 'good' if they are have never been more than one month in arrears.

In the definition of bad 'currently' represents a cut-off point usually taken between 14 and 18 months after the loan term has commenced. The class of bad accounts is formed from those that fall three months in arrears before the cut-off and those that fall two months in arrears at the cut-off point.

These definitions lead to a set of indeterminate customers who are neither 'good' nor 'bad'. Information on such accounts is readily available yet these are often left out from the analysis completely.

From this point on we will adopt having ever been three months in arrears as the definition of bad. This removes any arbitrariness arising from the 'currently' through the 'two months in arrears' part of the original definition. Also 'good' will, unless explicitly stated, be taken as the complement of bad. Once again, those simplifications are to permit us to develop models. These models would need refinement to reflect the operational definitions if they were to be applied in practice.

This thesis addresses several sources of uncertainty and variability that arise in credit scoring. Already we have eliminated one source of uncertainty by removing the 'currently' component of the UPL definition given above. Domain experts justify this aspect of the definition on business grounds. However, the value that represents 'currently' may vary between unsecured personal loan scorecards.

Consider the 'three months' in arrears as stated in the definition. One could argue that this choice is just as arbitrary as the 'currently' component, which we have chosen to remove from the definition. Three months may be no more appropriate than two or four months. However, these sources of uncertainty are different. All scorecards constructed for unsecured personal loans at the sponsoring bank incorporate three months in arrears into their bad account definition whereas the time point that represents currently may vary. Consequently two scorecards using different values for 'currently' but the same number of months in arrears would be constructed from models built according to different classes. Three months is certainly arbitrary, but the arbitrariness incorporated into the models is always the same. Sources of definition arbitrariness of this kind are addressed in Chapter 5, and other sources of uncertainty are also discussed at length in subsequent sections.

Although there are many sources of variability and uncertainty present in credit scoring, the possibilities for the behaviour of an account are limited. All accounts in the data sets can only adopt one of the following patterns of behaviour:

- Pay on time throughout the lifetime of the loan to fulfil the loan agreement.
- Default and fall into arrears.
- Default completely and be written-off as a bad debt by the bank.
- Decide to pay-off the remaining balance of the loan ahead of schedule.
- Be censored with loan term longer than the observation period.

3.2 Modelling Techniques

This chapter will explore the UPL(1) data set using techniques in common use in the credit industry. Our modified definition regards a customer as bad if that customer's loan account has fallen 3 or more months into arrears. In this section we will use this definition to replicate the process of constructing a scorecard in industry. Any results obtained should be similar to those which would be obtained and consequently implement in a commercial environment.

Using a definition which may legitimately be used in industry as a starting point is preferable to using a completely artificial definition because meaningful comparisons may be made. This would also aid implementation of any new techniques thought to be worth exploring in a commercial environment.

The history of credit scoring (Lewis, 1992a) shows that attention has been focused on improving existing scorecards in an attempt to construct the best classification rule according to some criterion such as error rate or Gini coefficient. Results in this chapter will discuss two methods in widespread use in the credit industry. Subsequent chapters will concentrate on the formulation of the problem and how to use the available data to produce results that may provide more reliable predictions. We believe this course of action more likely to result in increases in performance than attempting to construct an increasingly complicated classification rule.

In Chapter 2 we have seen that many of the variables used in scorecard construction are categorical. Indicator variables are frequently used in credit scoring. Sometimes these arise naturally from application questions such as 'does the applicant...', hold a cheque guarantee card, require credit protection insurance or have a telephone. Variables with an underlying continuum, such as age, time at address, or time in employment are grouped coarsely and entered into the analysis as two or three indicators rather than a continuous variable. Other categorical variables without an underlying trend such as credit cards held, occupation, loan purpose are grouped together if considered to exhibit similar patterns. In this thesis we represent each level of categorical variables by constructing indicator variables. This was thought appropriate because, as stated above, many variables used in credit scoring are already in this form. Additionally, any methods of analysis we propose could readily be adapted to deal with other variable transformations.

When considering a variable with n levels one problem associated with using indicator variables is that (n-1) indicator variables must be incorporated into the analysis. The dimensionality of the data set can increase very quickly in these circumstances. As n gets larger the risk of overfitting is increased.

Methods are available to transform qualitative variables into quantitative variables see Crook, Hamilton and Thomas 1992 and Boyle et al. (1992) for details. One method in common use in the credit industry is *weights of evidence*, which are defined as:

$$w_{ij} = \ln(G_{ij} / B_{ij})$$

Where G_{ij} is defined as the number of goods in the jth level of variable i divided by the total number of goods and B_{ij} is the number of bads in the jth level of variable i divided by the total number of bads. These weights of evidence replace the actual predictor variable values for analysis.

The variables used in the following analyses are similar to those used in the construction of unsecured personal loan scorecards in industry. Logistic regression is a common approach. A combination of domain knowledge and stepwise variable selection methods is often used in order to optimise the discriminatory ability of logistic regression. Other techniques may have produced superior results using an alternative subset of variables. Moreover, because we use a variable set that performs well with logistic regression, any observed improvement in performance using other techniques will be conservative. In the absence of domain knowledge about the vast number of other variables available it was considered a sensible approach to adopt a variable set similar to that used by the sponsoring bank.

The UPL(1) data set consisted of 22490 observations after anomalous observations were removed (Section 2.3.3). As mentioned in Chapter 1, when constructing a scorecard to be used in the credit marketplace it is standard practice to build the model using a subset of the data and validate the model using the remainder of the data to assess the performance of the model. This provides an estimate of future performance that is independent of the design set. Here these data were randomly split into a design set consisting of 15595 observations (70% of the original data set) and a test set of 6895 observations (30%).

3.3 Linear and Logistic Regression

The two classification techniques which are most widely used in credit scoring are linear and logistic regression, (Henley, 1995). Both methods lead naturally to a score table. This is favoured by financial institutions because a reasonable justification can easily be given to explain why an application is rejected.

3.3.1 Linear Regression and Discriminant Analysis

It has been shown numerous times that linear regression and discriminant analysis for the two class case are equivalent. Regressing Y on $\mathbf{x}$, where Y is the

response variable and **x** a feature vector gives the same linear function as Fisher's discriminant analysis. For examples see Leonard (1988), Orgler (1971).

The popularity of linear regression in the credit scoring community is most likely due to its conceptual simplicity and the ease with which a solution can be computed. A linear model which relates the response vector $\boldsymbol{y}$ to a set of independent variables $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_k$ takes the form:

$$y_j = \beta_0 + \beta_1 x_{1j} + \beta_2 x_{2j} + \ldots + \beta_k x_{kj} + \varepsilon$$

where y=1 if the observation is from the bad class and y=0 otherwise, and $\varepsilon \sim N(0, \sigma^2)$. Solutions can easily be estimated using least squares.

However, despite the widespread use of linear regression, is not considered a theoretically sound basis upon which to implement scorecards (Henley 1995, Sewart 1997).

When performing linear regression the response variable is assumed to be normally distributed and variance of the observations constant. Modelling a binary response violates the constant variance assumption. However, these assumptions are required for inference on the model parameters and reliable significance testing. We are concerned only with the predictive power of the model. Reichert, Cho and Wagner (1983) discuss issues involved in building scoring models. Their opinion is that the predictive power of linear regression is not impaired by failing to meet the normality assumptions. Indeed, in our experience many techniques including linear regression give much the same classification results (see Figure 3.1), even though other techniques, such as logistic regression, appear more theoretically sound.

3.3.2 Logistic Regression

Logistic regression is one technique which is more theoretically suited to modelling binary response data than linear regression. The logistic regression model is given by:

$$\text{logit}(p) = \log\left(\frac{p(1\,|\,\mathbf{x})}{p(0\,|\,\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_k x_k$$

Where $\boldsymbol{\beta} = (\beta_0, \beta_1, \ldots, \beta_k)$ is a parameter vector.

An advantage of logistic regression over linear regression is that a simple transformation provides predicted probabilities of class membership. Predictions from linear regression are not constrained between 0 and 1and so are not true probabilities. Logistic regression is a special case from the class of generalised linear models (GLMs). For an authoritative discussion of GLMs see McCullagh and Nelder (1983).

Titterington (1992) highlights difficulties associated with logistic regression. Firstly, the maximum likelihood estimates need to be calculated by iterative numerical algorithms. Modern computing power has successfully addressed this problem. Titterington also points out that logistic regression cannot easily be implemented when the response is assumed to lie on a continuum of creditworthiness. This point relates to a response that has many classes and is redundant when dealing with the two class problem, as is often the case in credit scoring. However, more standard software packages are now equipped with multi-class logistic regression routines which partially addresses this problem. In addition, Chapter 5 discusses an alternative approach to addressing credit scoring problems when the classes arise from a continuum.

Figure 3.1 shows the resulting ROC curves obtained from linear and logistic regression classification rules. Each rule was formulated using fourteen variables of which eleven were entered into the analysis as indicator variables.

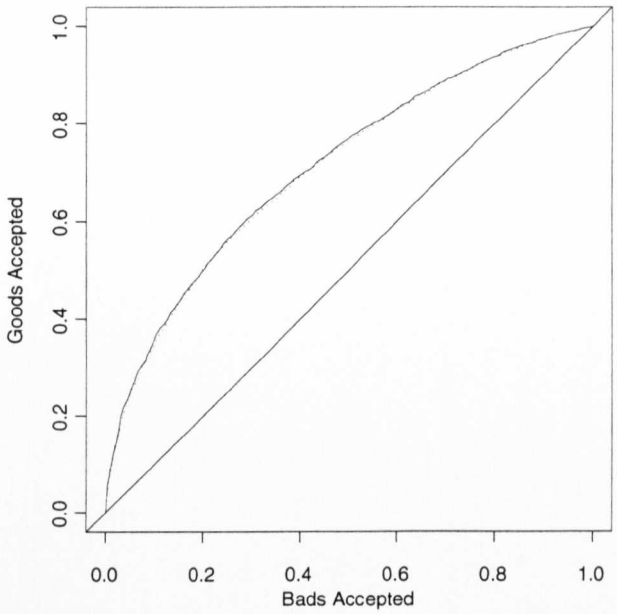

Figure 3.1: *ROC Curves for Logistic and Linear regression*

Figure 3.1 Shows the ROC curves obtained from the logistic classifier (solid line) and classification rule based on linear regression (dotted line). Performance is almost indistinguishable from the figure alone. The Gini coefficients obtained are 0.4274 and 0.4197 for logistic and linear rules repectively.

Chapter 4

Possible Improvements

4.1 Introduction

Chapter 3 outlined practices commonly used throughout the application scoring sector of the credit industry. In this chapter we discuss possible improvements in the form of:

- **An alternative modelling technique** – Survival analysis will be used to incorporate a time aspect into the scorecard.
- **Using alternative definitions to represent a bad account** – Various alternative definitions are proposed and modelled.
- **Modelling profitability** – A measure of profitability is proposed and modelled.

Throughout this chapter we will provide examples using the UPL(1) data. This is because the size of UPL(1) is typical of a data set used for commercial scorecard construction.

In later chapters we will further extend these ideas to produce novel ways of tackling some of the problems occurring frequently in credit scoring.

4.2 Survival Analysis

Similarities may be drawn between the frameworks of an account falling into arrears and disease recurrence in medical contexts. An account may be conducted well until a payment is missed whereas a disease may be in remission until recurrence is discovered. There is a large literature concerned with survival analysis. Cox and Oakes (1984) and Kalbfleisch and Prentice (1980) provide texts with a good depth of the theoretical concepts of survival analysis. Collett (1991) and Parmar and Machin (1995) give practically oriented descriptions. Much of the survival analysis methodology was motivated by medical and engineering problems. In the financial environment many outcomes could be modelled. For example, missing a payment, write-off or early settlement. Narain (1992) noted that for accounts with different characteristic profiles, the outcome variable can be modelled as a survival function. He performed some basic analysis that indicated survival techniques may be potentially valuable in the credit area. However, little has been published since. Banasik, Crook and Thomas (1998) provide a detailed comparison of survival analysis and logistic regression. The authors investigate proportional hazards models, accelerated life models and a competing risks approach. They conclude that further research on survival analysis could prove useful in credit scoring problems. Suggested routes are to allow time dependent features in the models, or use distributions other than the standard Weibull as a basis for modelling. We demonstrate that proportional hazards survival analysis (Cox, 1972) can yield performance comparable to traditional credit scoring techniques.

Many events of the form 'time to .....' are of potential interest in credit scoring. For example, if time to bad could be modelled accurately then accounts that go bad at a late stage, but still yield an overall profit can be accepted. At the same time accounts that go bad very quickly and represent a substantial loss for the bank can be minimised. Outcomes of this type could be modelled using survival analysis techniques.

Problems arising from the UPL data can be classified into two distinct types:

1. Behaviour over time can be predicted using the application data alone – application scoring. Essentially, current techniques model this framework. Outcomes such as 'time to default' or 'time to early settlement' could be modelled.

2. In addition to the application data alone, the behavioural data, up to a certain point in time, can be utilised to model future behaviour. Outcomes such as 'time to pay-off given x months default' or 'write-off given attainment of some lesser state of arrears' would be appropriate for investigation.

4.2.1 Cox's Proportional Hazards

In credit scoring the hazard function is the probability of failure (an account becoming bad) in the next small time interval, given that the account has remained good to the beginning of the interval, and is defined as,

$$h(t) = \lim_{\Delta t \to 0} \frac{P(t \leq T < \Delta t \mid T > t)}{\Delta t},$$

with T a lifetime random variable.

Our aim is to relate the survival time of accounts to different feature vectors. Cox's proportional hazards model is a semi-parametric model which assumes that accounts have the same shape of hazard function but shifted multiplicatively. The hazard rate of any particular group is assumed proportional over time. Other than this, no other assumptions are made about the underlying hazards. One advantage of the proportional hazards model is that it allows a non-constant hazard rate to be modelled and it is not important to know the actual distribution of the failure times as these are not parametrically specified.

The general proportional hazards model for the mth account at time t is given by:

$$h_m(t) = \exp(\beta_1 x_{1m} + \beta_2 x_{2m} + \ldots + \beta_k x_{km}) h_o(t)$$

Where $x_{1m}, x_{2m}, \ldots, x_{km}$ are the predictor variables and the baseline hazard for the 'average' customer is defined as $h_0(t)$.

This is often written as the logarithm of the hazard ratio because this can be regarded as a linear model:

$$\log\left\{\frac{h_m(t)}{h_o(t)}\right\} = \beta_1 x_{1m} + \beta_2 x_{2m} + \ldots + \beta_k x_{km}$$

Survival analysis techniques are often used in medicine to predict the time to a particular event such as disease recurrence or pregnancy, or in engineering to predict the lifetime of an electronic component. We intend to use the methodology to classify accounts on the basis of their likely survival time.

4.2.2 Scope and Use of the Model

The effect a single binary variable has on the time taken for an account to be classified (by the standard definition) can be modelled. Figure 4.1 is an artificial example that shows a marked difference between those accounts with the variable present and those without.

Figure 4.1 shows the survivor functions of two accounts, the solid line, an account whose profile included the variable and the broken line an account that did not. Accounts where the variable was not present tend to fall three months in arrears at a faster rate.

Each account has a variable profile, that is, an associated feature vector. As the model is constructed using a great number of variables the number of possible different feature vectors increases rapidly. For example if ten binary variables are included in the model then there are 1024 distinct profiles. Graphical

representations of the survivor functions for different profiles become difficult to interpret.

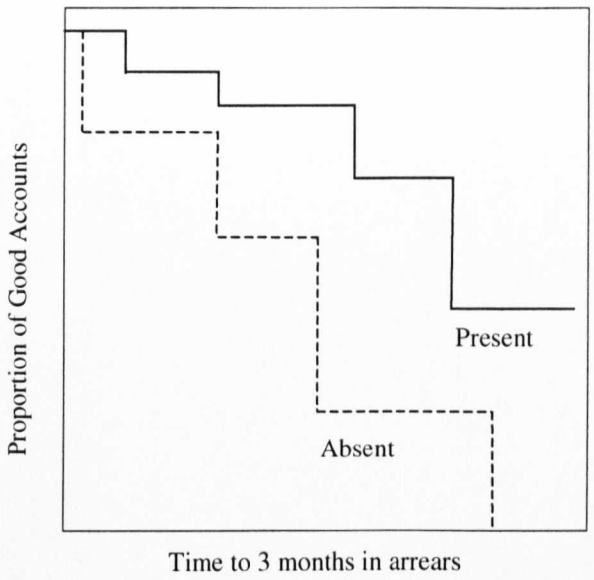

Figure 4.1: *The effect on survival time: a univariate example.*

When considering accounts with a high potential for default it is likely that interest will be directed at accounts defaulting before a certain time, t, of their repayment schedule. This t may vary depending on many factors such as loan term and loan amount. Consider 36 month loans. Suppose we knew that, on average, 36 month accounts that are classed as bad before t monthly repayments incur a loss. In this case a model could be constructed to predict those accounts most likely to turn bad before this point. Survival analysis provides an appropriate class of models. This information could be used to impose restrictions on the granting of loans to customers with high propensity of falling bad early in the lifetime of the loan agreement.

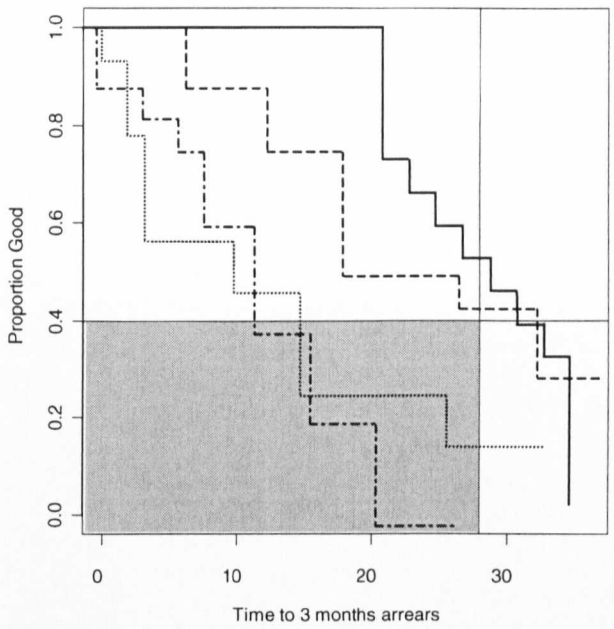

Figure 4.2: *Illustration of using proportional hazards to classify on the basis of likely survival time.*

When considering acceptance policy a bank may choose to accept applications on the basis of the proportion of customers, p, with feature vector $\mathbf{x}$, who are predicted to be good after t months have elapsed. Figure 4.2 illustrates when p=0.6 and t=28. Credit will not be granted to customers whose profiles fall in the shaded region.

4.2.3 Comparison of Proportional Hazards Regression and Logistic Regression

Initially, the idea of a survival analysis model was to enable predictions of the time at which an account would fall bad. However, we found that predictions for the time at which an account actually turns bad are not sufficiently accurate to be of commercial use. On the other hand, the survival models could be used to predict which accounts would become bad before a certain time, t. By imposing this threshold t we have reduced the proportional hazards model to a dichotomous

outcome – an account is either bad before t months or it is not. No longer would we have a prediction of the actual time of turning bad. This definition could be modelled by logistic regression techniques simply by incorporating t into the definition at the outset. Figures 4.3(a) and (b) show that a proportional hazards model performs as well as logistic regression (solid line) with t=12 and t=24 months.

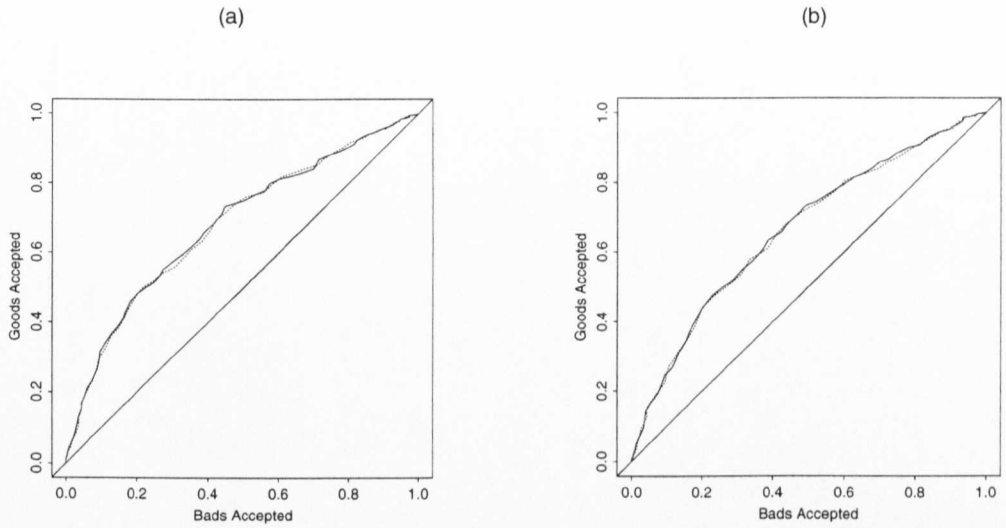

Figure 4.3: *Comparison between logistic and proportional hazards models for accounts which turn bad before 12 (a) and 24 months (b).*

Two important points arise from this analysis. First, the predictions of the time at which an event takes place are not sufficiently accurate to facilitate reliable classification. Second, there is no substantial improvement over constructing a classification rule via modelling a binary variable. We conclude that proportional hazards provides a viable alternative to logistic regression, yet offers no considerable improvements which make it particularly appealing.

4.3 Experimenting with Different Definitions

Scorecard splitting was briefly introduced in Section 1.1.5.5. When modelling unsecured personal loans, scorecards are often split by age. This procedure involves building a different scorecard for each of several different age groups. As noted earlier, splitting has the same effect as including the interactions

between the splitting variable and the other variables in the scorecard. Improvements resulting from scorecard splits using variables such as age are often small. Age is an intrinsic property of an account holder. The predictive power of age can be can be explained solely by the differences across subgroups in the population. However, other variables produce fundamentally different patterns of repayment that are not caused by the population. Variables of this type impose deterministic structure on the data that is caused by extrinsic properties such as loan term. Consequently, larger differences between the split scorecards may be realised when basing scorecards on this latter type of variable. The deterministic structure imposed by variables such as loan term is comparable to that which would emerge if certain credit products were only available to different age groups.

Consider two randomly chosen accounts. If the ages of the account holders differ, then one account may have higher potential of falling into arrears. However, considering two randomly chosen accounts whose loan terms are different produces different repayment behaviour. Obviously the account with longer loan term has a greater length of time to default, with more repayments that could be missed. Moreover, the economic climate could substantially change during the period of time not common to both accounts, thus introducing further sources of variability. Simply treating accounts collectively and including the splitting variable as a predictor variable will benefit the scorecard. However, due to the underlying deterministic structure, scorecard splits based on variables such as loan term are likely to benefit scorecard performance. In Section 4.3.1 we treat accounts with different loan terms distinctly and assess the effect this has on classifier performance.

4.3.1 Loan Term and Definitions of Badness

Financial institutions often treat their whole database of unsecured personal loans collectively and apply the same class definitions throughout. However, different loan terms produce fundamentally different patterns of repayment. A twelve

month loan has very little time to go bad. To fall three months in arrears (the standard definition) requires a twelve month loan account holder to default on 25% of repayments. The same definition applied to sixty month loans requires default on only 5% of repayments. Interesting patterns in the data, which may have been concealed when all loans are treated equally, may be investigated when addressing the problem from this perspective.

Examination of the data set shows that the percentage of bad customers strictly increases as the loan term increases (illustrated in Table 4.1).

Loan Term (months)		Standard Definition % Bads
12	(25.00%)	7.9
24	(12.50%)	13.2
36	(8.30%)	18.4
48	(6.25%)	23.0
60	(5.00%)	24.3

Table 4.1: *Breakdown of Good/Bad by standard definition with respect to Loan Term.*

The bracketed figures after loan term in Table 4.1 indicate the percentage of repayments which must be defaulted upon in order for an account to be classified as bad by the standard definition.

The standard approach assumes that each customer has a probability, p, of becoming bad at time t. Consequently, one would expect longer loan terms to have higher proportions of bads. This approach is appropriate when the outcome being modelled is unquestionable. However, in this case we are modelling creditworthiness, which does not have a precise definition. Previously we noted that different loan terms necessarily produce different repayment behaviour. Accounts with different loan terms and loan amounts require different numbers of repayments before the account yields profit. Using the three months in arrears definition on different accounts fails to segment the population into fundamentally distinct categories. A twelve month account holder classified as bad would be regarded as worse than a five year account that also missed three monthly payments. The former would be considered a serious defaulter whereas

the latter, a forgetful payer. Given the varying difficulty in which accounts with different loan terms can be classed as bad it may be preferable to use an alternative definition of bad. The definitions proposed below ensure that 'bad' is more similar between different levels of loan term.

Taking into account the varying loan terms we propose the alternative definition:

$$\text{Bad} = \frac{\text{Loan term}}{12} \text{ months in arrears,} \qquad [1]$$

Using Definition [1] we obtain Table 4.2.

Loan Term	Alternative Definition % Bads
12 (1)	32.0
12 (2)	13.3
24 (2)	20.2
36 (3)	18.4
48 (4)	18.5
60 (5)	17.3

Table 4.2: *Breakdown of Good/Bad by using Bad =(Loan Term)/12 months in arrears as the Definition of badness with respect to Loan Term.*

In Table 4.2 the figures in brackets after loan term indicate the number of months of consecutive defaults required for an account to be classed as bad under the new definition. The new definition for the 12 month loan term leads to a very large percentage of bads. This is perhaps too strict a definition, since an overlooked payment rather than meaningful default can easily explain one month of arrears. Definition [1] can be modified to give:

$$\text{Bad} = \frac{\text{Loan term}}{12} \{+1 \text{ if loan term is 12 months}\} \qquad [2]$$

Although ad hoc, this modification means that the percentage of bad accounts is maintained at a broadly similar level throughout loan terms. There is some degree of arbitrariness associated with all bad definitions. However, even with

modification, Definition [2] introduces a mechanism (i.e. proportion of bads) by which different loan terms may be compared.

Figures 4.4(a) and (b) show five ROC curves which correspond respectively to the different loan terms for the standard definition and Definition [2]. The classification rules were constructed using logistic regression.

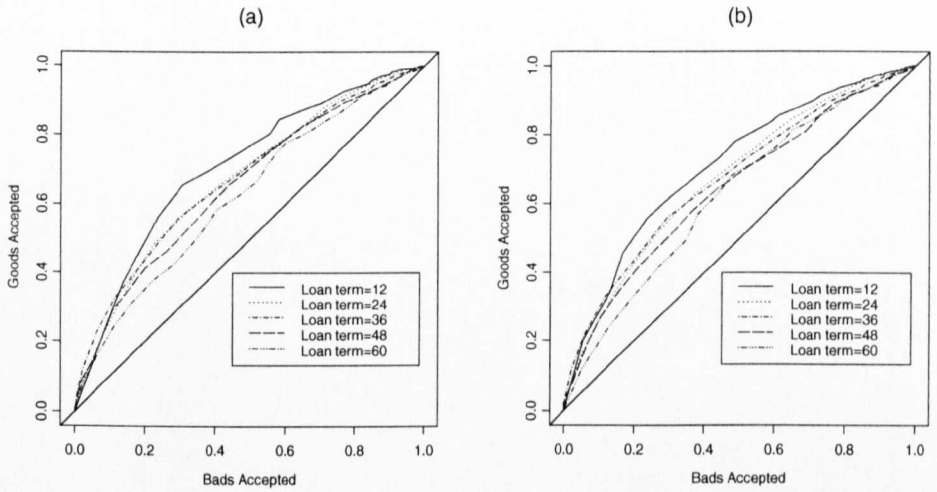

Figure 4.4: *ROC curves to show how the standard definition (a) and Definition [2] (b) perform with respect to loan term*

Figure 4.4 clearly shows the classification rule for 12 month loans performs considerably better than all the other loan terms. In addition, the general trend is of decreasing performance, as illustrated in Table 4.3. The performance of the classification rules constructed for 48 and 36 month loans are similar.

Each of the classifiers obtained from the alternative definition perform very similarly to those from the original definition. This is somewhat surprising since Definition [2] was formulated on the basis of the proportion of bads at each level of loan term. Table 4.3 shows that the alternative definition performs worse only on the 48 month loans, the 36 month loan term has the same performance for each approach.

Loan Term	Gini coefficient Standard Def	Gini coefficient Definition [2]
12	0.415	0.419
24	0.344	0.349
36	0.346	0.346
48	0.307	0.288
60	0.245	0.245

Table 4.3: *Gini performance of the standard definition compared to Definition[2].*

Accounts with 36, 48 and 60 month loan terms are essentially the same for the first 36 months. The lengths of their remaining payment periods vary. An alternative definition would be one that would produce the same bad rate across different loan terms. We consider this to have advantages because implementing different definitions such that the proportion of bads is similar across different loan terms presumes that there is an underlying proportion of the population that are bad customers. Whereas applying the same definition across different loan terms causes large differences in the proportion of bads defined in the sub-populations of loan term.

We have shown that classification rules built according to different definitions have very similar performance, yet the performance can vary substantially when compared across different sub-populations. Interesting questions arise when this is the case. For example, would a customer applying for a three year loan be rejected, yet accepted had they applied for a five year loan? Scallan (1997) discusses a scoring system which could determine whether an unsuccessful application may have been granted a loan of a different term.

The scorecards constructed using linear and logistic regression in Chapter 3 assign lower scores for longer loan terms. This reflects the increased propensity for a customer with a longer loan term to fall into the bad category. If the bad definition used is based on the number of months in arrears, then the very nature of a five year loan implies that more accounts will be bad in comparison with shorter loan terms. Figure 4.4(a) shows that the classification system based on the

standard definition performs better, in terms of Gini coefficient, on shorter loan terms.

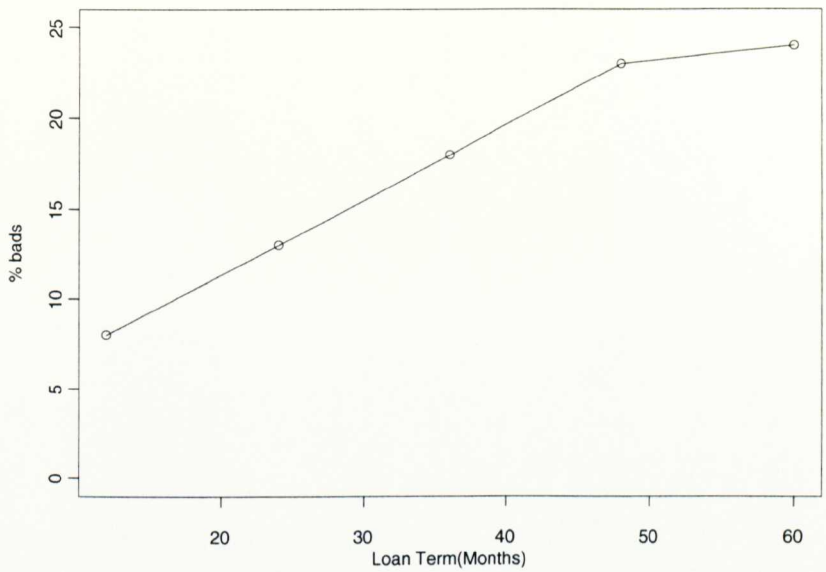

Figure 4.5: *Rate of Increase of Percentage Bads with respect to Loan Term.*

Figure 4.5 shows an approximate linear relationship between percentage of bads and loan term. The percentage of bads when the same definition is applied will increase simply because, 'the longer the loan term, the greater the opportunity to become bad'. This data set only has three years of data for four and five year loans. The proportion of bads (standard definition) is increasing with increasing loan term as shown by the larger proportion present in 12, 24 and 36 month loans. In addition the rate at which the bad accounts accumulate is accelerating. During the three year observation period larger proportions of bads are accumulating for loan terms longer than three years than the proportion of bads present in three year loans. This phenomenon is illustrated in Figure 4.6(a).

The above results may lead to the claim that a higher proportion of customers become bad, at a faster rate, when considering loans of longer term. From the business viewpoint this is the exact opposite to the desired effect. If accounts

with longer loan terms are becoming bad quickly then only a small proportion of the loan capital has been repaid, along with none of the accounts' potential profit. These policies are effectively accepting more of the largest loss making customers.

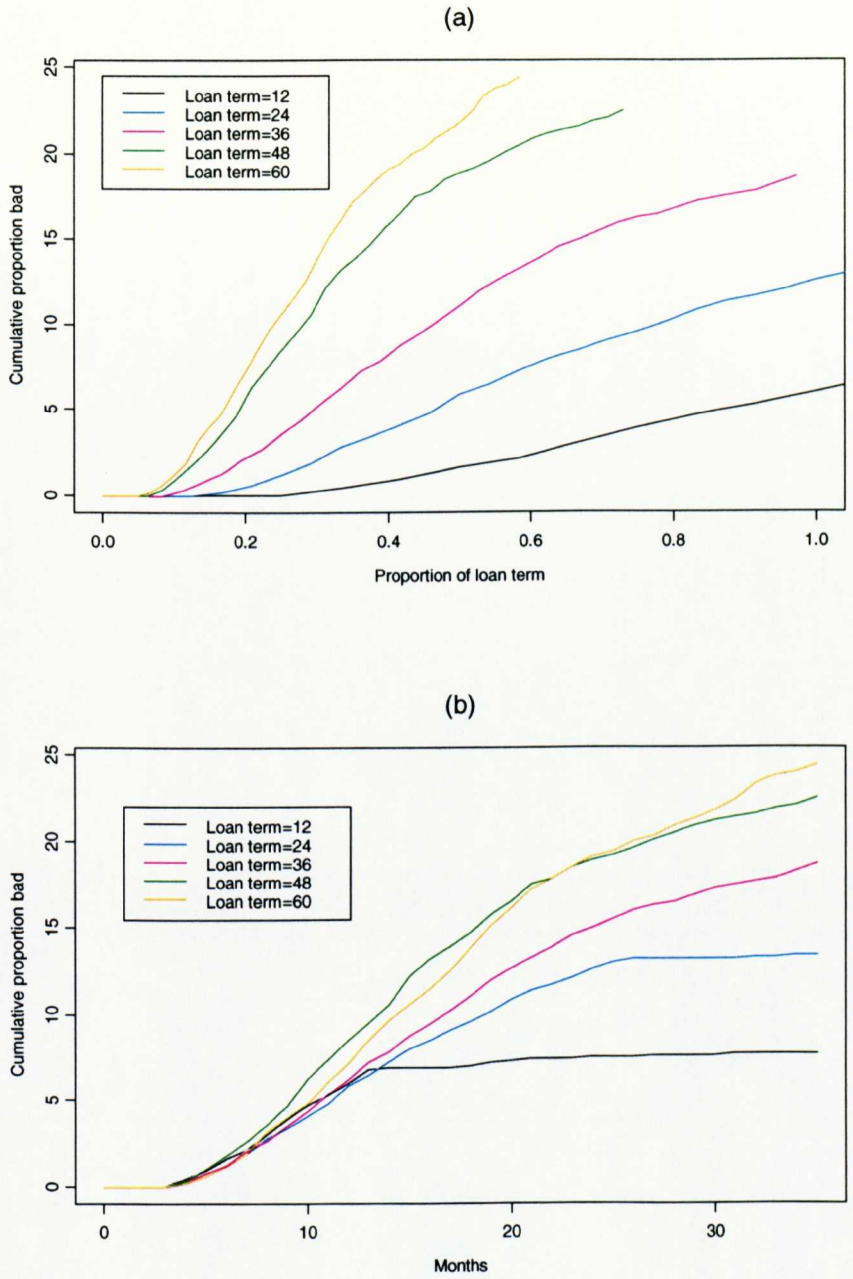

Figure 4.6: *The cumulative bad rate for different loan terms (standard definition).*

When considering Definition [2] it is apparent that the longer the loan term, the more extreme the definition of bad becomes. Let us consider for example 60 month loans. Using five months in arrears as the definition of bad substantially reduces the number of bads in the population compared with using the standard definition. However, this is during the first three years of the loan. There is a large percentage of customers getting heavily into arrears rather quickly.

Figure 4.6(b) shows that the different loan terms begin to separate noticeably after about 12 months. At 24 and 36 months the bad rates are substantially larger for the longer loan terms.

Were the entire data available, it would be reasonable to assume that the percentage of bads would be higher than that observed for 48 and 60 month loan terms. Under the standard definition bad rates for these two loan terms are already approaching 25%, so surely regarding such high values as 'normal proportions of bads' only reinforces the view that the current definitions are not an appropriate means of addressing the issues central to the problems associated with credit scoring. This is due to the absence of a consistent definition of bad.

4.3.2 Predicting Different Outcomes

Rather than attempting to find a better model for the standard three months definition, perhaps the problem should be re-evaluated. An alternative definition may permit better predictions of future performance. If the alternative is appropriate then it may be preferred. Alternatively a definition may be geared towards predicting profit rather than default of some sort. Perhaps one of the most crucial issues in credit scoring is whether an account makes a profit or a loss. Ultimately, if a loan account is repaid entirely then the number of missed payments is irrelevant. Missing monthly payments can result in extra interest penalties. Consequently, this type of account may be the largest profit makers. Moreover, the bank should not necessarily be interested in identifying customers

who fall some number of months in arrears, but should rather be interested in identifying accounts which make a loss.

Many definitions may be investigated. Numerous simple extensions to the current definition are possible, for example:

- Whether a customer falls a different number of months into arrears.
- The total number of missed payments during the lifetime of the loan.

In this chapter we have already suggested incorporating time into the problem. Time can be incorporated in two ways. Firstly, by modelling the time to an event such as write-off or early settlement. Secondly, as proposed in Section 4.3.1, by incorporating loan term into the definition.

There are many potential outcome definitions: four months in arrears, three months–with two of the arrears arising from consecutive months, bad in the first 18 months, all represent sensible alternative definitions which may be modelled. Several of these models were tested – none provide vastly superior results to the model based on the standard definition.

Figure 4.7 show the ROC curves derived from using logistic regression to model the following definitions:

1. The standard definition.
2. Accounts classed as bad by the standard definition during the first twelve months.
3. Accounts written-off as bad debt.

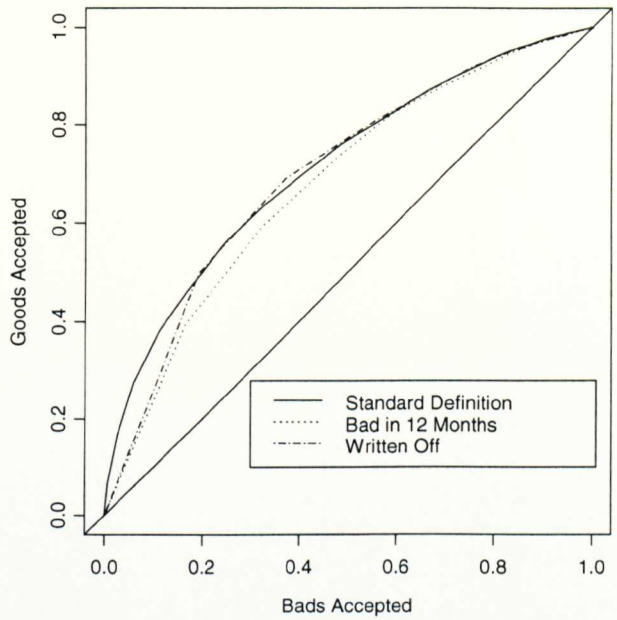

Figure 4.7: *ROC Curves for different definitions of bad.*

ROC curve analysis of the classifiers built for these alternative definitions produced very similar results for the majority of definitions. Perhaps there is a common underlying factor that could be thought of as the core component of a bad account. The account being written-off may represent such a common factor.

Suggesting an alternative in this manner is just as arbitrary as the original definition. Chapter 5 addresses the choice of definition by proposing models that allow the definition of bad to be altered once the model has been constructed.

Adopting a definition based on customers who settle early is one of the few outcomes that gave markedly different results when compared with the standard definition. Modelling early settlement using the whole data set leads to a classification rule with poor discriminatory power (Figure 4.8).

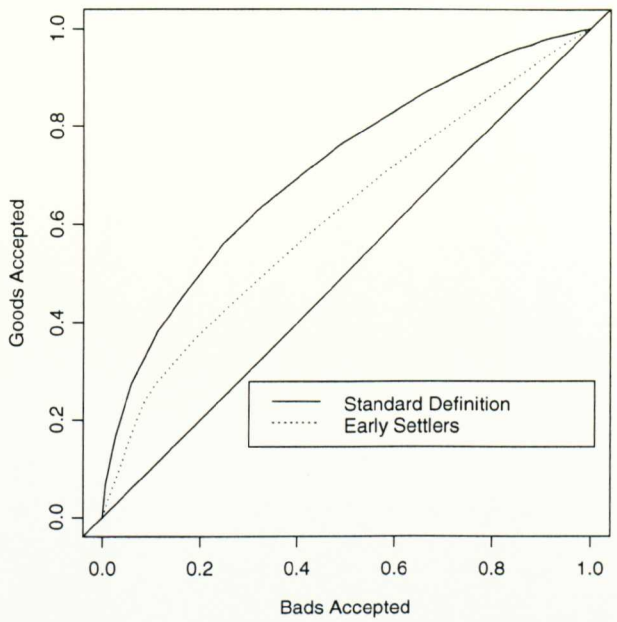

Figure 4.8: *ROC Curves for the standard definition of bad and early settlers.*

The performance of the classification rule for early settlement is extremely poor, so it may be of limited practical use. However, note that an incorrect decision when predicting early settlement is not so costly. Modelling good and bad accounts at the credit granting stage may result in undesirable losses when incorrectly classifying bad accounts as good. To recoup the losses associated with a misclassification of this type the bank may require up to ten good accounts. If we were to model early settlement amongst the accepted population, lost revenue is lost profit rather than a monetary loss. Consequently, a classification rule for early settlement may be useful in addition to the standard scorecard and be used to assign an early settlement clause in the loan agreements of those accounts predicted to settle early.

4.4 Profitability

Performance may be defined in different ways. Actual amount outstanding at the end of the loan term may be used as a measure of creditworthiness. On the basis of a 12% interest rate, loans of £750 and £5000 would have respective total amounts payable of £840 and £5600. These would be considered to have performed equally if each had £200 outstanding at the end of the loan period. The bank may regard a £200 loss as extremely undesirable, and perhaps work on the basis of a fixed potential loss per account. On average this strategy may lead to overall success. However, this might not seem ideal. £200 is a much larger proportion of the first loan than the second. Much of the capital of the small loan has been lost to bad debt whereas all the capital and some of the interest on the second loan was repaid prior to the account being classed as bad. An alternative performance measure, which allows for this, is the amount outstanding at the end of the loan period expressed as a proportion of the total interest payable on the loan. This measure assesses the quality of account conduct by relating the actual amount outstanding to the total loan amount.

Adopting this approach, the proportion of interest outstanding for a perfectly conducted account should be 0. Those accounts with proportion of outstanding interest greater than 1 are far less desirable. At a time when the whole loan should have been repaid they have not even finished repaying the capital. Values between 0 and 1 represent a sliding scale of how much of the interest is still outstanding. The proportion of interest outstanding that is to be regarded as acceptable can be decided. A definition based on this proportion can then be used to model this profitability index.

This profitability index does not give an indication of the amount outstanding yet is comparable across different loan amounts. In the example given above, one loan has not even repaid the capital (222% of the interest still unpaid), whereas the other has already repaid two thirds of the account's interest (33% of the total interest payable still owing).

4.4.1 An Example

As we have seen in Section 4.3, a proxy such as months in arrears is often used for profitability. Consider months in arrears. If a loan has term n and the bad definition implemented is x months in arrears then x/n is the proportion of the total loan amount that is considered to be sufficiently large that the bank would reject any application predicted to have problems repaying that amount.

This is similar to what we are proposing. However, the profitability measure described above would only accept customers with a certain proportion of interest outstanding. Let us consider two examples. First, a definition used to model accounts which are predicted to have paid at least 50% of their interest by the end of the loan term. Second, a relaxed definition which aims to predict accounts that have repaid 20% of their interest.

Constructing two classification rules based on logistic regression using these definitions gave the ROC curves in Figure 4.9.

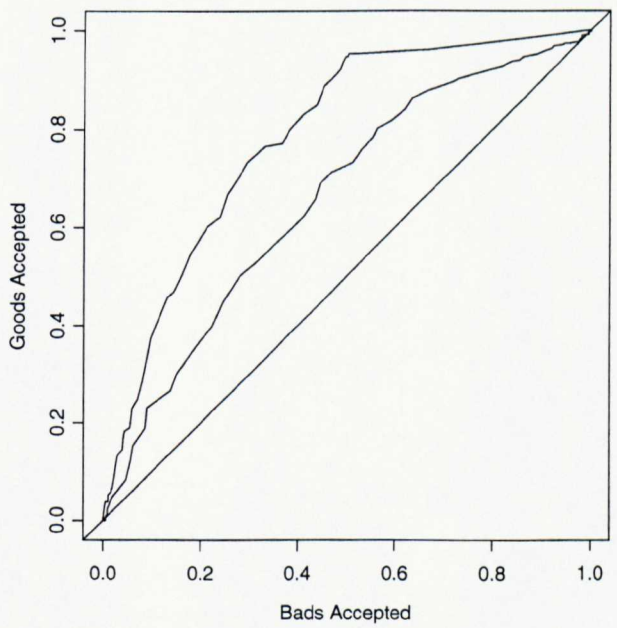

Figure 4.9: *ROC curves for 0.2 and 0.5 profitability thresholds.*

The larger ROC curve in Figure 4.9 represents the performance of the classifier constructed according to the 0.2 profitability index threshold. The smaller curve represents a threshold of 0.5. These curves have Gini coefficients of 0.568 and 0.313 respectively.

These results alone are not sufficient to suggest that such a definition would outperform the standard definition. Banking expertise would be required to determine a suitable definition of this form to use in industry. However, the interesting point here is that by changing the profitability definition, classification performance is greatly influenced. In the previous section it was often the case that, when dealing with different definitions, little change in classifier performance was evident.

4.5 Conclusions

In this chapter we have investigated the effects of changing class definitions on the quality of prediction. We have seen that deviations from, and alternatives to the standard definition do not result in notable changes in performance. However, by considering profitability we have shown that predictive power can be increased substantially. We have also seen that varying loan terms produce accounts whose repayment patterns are fundamentally different. Longer loans have a higher bad rate as there is more time for an account to become bad. It seems that accounts with longer loan terms become bad at a faster rate when using the standard definition. These points, and the results in Section 4.3, suggest that treating different loan terms as different cases, or redefining the definition of a bad account accordingly, may lead to improvements in scorecard performance.

In credit scoring, high misclassification rates are commonplace. Varying the definition of a 'bad customer' can produce classification rules that perform differently. Although these differences are often small, this is an industry in which a small difference in predictive power may imply a change in profit of millions of pounds. Subsequent chapters will investigate dynamic models that

allow definitions to be altered according to external conditions and methodically explore regions of possible definitions such that the most useful definition may be implemented.

Chapter 5

Global Models

5.1 Introduction

Typically in credit scoring class definitions partition the population into good and bad classes. The definitions represent some notion of creditworthiness. A scorecard is constructed by building a predictive model that discriminates between the classes. Chapters 3 and 4 discuss various definitions of bad unsecured personal loan accounts.

A scorecard is built and then used in the credit marketplace for a limited period of time. Statistical modelling attempts to predict which customers are most likely to be classed as bad, so that such customers can be declined when applying for credit. While the scorecard is in use, external influences such as economic factors, are continually changing. This tends to result in the classification performance of the scorecard deteriorating. Classifier deterioration can prove costly when considering the profit that would have been gained from good accounts that were inadvertently rejected due to the inferior performance of the scorecard. In addition, a financial loss is demonstrated by awarding credit to bad applications which, had the scorecard been discriminating with the same degree of accuracy as when it was initially implemented, would have been rejected. Usually when a scorecard is deemed to have deteriorated sufficiently a new one is built. This remodelling process can be expensive, both in terms of time and

money invested in the construction process required to build a new scorecard and so should be avoided unless absolutely necessary.

The definitions of the 'good' and 'bad' classes are usually based on banking knowledge rather than formal analysis. Consequently there is arbitrariness associated with those definitions. The standard definition of bad for an unsecured personal loan is when an account falls three or more months into arrears. Months in arrears is used as a proxy for the profitability of an account. In the limiting case accounts with substantial arrears are likely to represent the greatest financial losses. However, as discussed in Chapter 1 the converse is not true. Those with least arrears are not necessarily the most profitable accounts. Three months may be chosen in the belief that the overall profit for the group of accounts not falling three months into arrears is greater than the profit obtainable from any other group specified by a different definition. However, this is not necessarily the case. Even if the definition, in the first instance, is correctly chosen to be the one that isolates the most profitable subgroup of accounts, there is nothing to say that this 'best' definition will not be affected by some external factor. It is quite reasonable to assume that for some non-zero proportion of time a scorecard is in use, an alternative definition may yield superior results. In the case of the unsecured personal loan example, 2 months or 4 months in arrears may represent definitions which may be preferable during some the operational lifetime of a scorecard.

Sections 4.3 and 4.4 proposed alternative ways of defining bad accounts. Relating the bad definition to the loan term or using a profitability measure were suggested and discussed. These approaches provide plausible alternatives that may have been used to formulate a bad definition. However, the partitioning value of the definition variable is still chosen in a somewhat arbitrary manner.

In this chapter we introduce models which do not require an exact formulation of a bad definition prior to modelling. Instead these models allow the precise choice of definition to be determined when the classifications are required. This enables

many definitions to be investigated or definitions altered without the need for remodelling.

5.2 Definition Threshold

Consider an unsecured personal loan scorecard. The scorecard may have been based on the definition '3 or more months in arrears' representing a bad account. This scorecard may have been performing well in the marketplace until company policy changed to dictate that the customer base of unsecured personal loan accounts should be increased. Alternatively, competitors may be using marketing strategies which results in business being drawn elsewhere. If this lost business is to be recovered one solution would be to change the definition which underpins the customer selection mechanism.

In Chapter 1 we defined a classification threshold which is the value that determines which class an observation with predicted probability $\hat{p}(1|\mathbf{x})$ is assigned. Here we introduce the idea of a *definition threshold.*

The definition creates classes by partitioning some continuum of creditworthiness, such as number of months in arrears or extent over overdraft limit. Hand, Oliver and Lunn (1997) discuss discriminant analysis when the classes arise from a continuum. We call the point on the continuum a definition threshold. Typically this definition threshold is chosen before the model is built, so the two classes are defined without scope for change. If circumstances changed such that an alternative class definition – an alternative definition threshold – were a more desirable definition to model, then it will normally be necessary to construct a new classifier based on this new definition. The alternative would be to incur the potential loss consequent on using the old classifier.

Remodelling every time the definition threshold needs changing is highly unsatisfactory. A better solution would be to find a *global model* which enables a

whole range of class definitions to be incorporated into a single model and where a decision on the class definition being modelled need not be made until the actual time of classification. This would eliminate the need for building a new scorecard each time the classification performance of the scorecard degrades and allow the bank to change policy freely with respect to bad accounts.

The ideas underpinning these global models could be applied to many areas such as marketing or behavioural scoring. In marketing, the campaign most likely to be successful could be determined by assessing the predicted probabilities of customers taking out a further loan when different incentives were offered. An application of behavioural scoring could be applied to a collections department. Statistical predictions for future months would enable debt recovery strategies to be initialised earlier than may otherwise have been possible. These predictions could each be assessed using a single global model simply by varying the definition threshold.

In this chapter we introduce a class of such models. We illustrate the benefits of these global models using various real and simulated data sets when considering classification of two classes arising from partitioning a single underlying continuum.

5.3 Global Models

A typical scorecard is constructed according to a definition which partitions some continuum of creditworthiness. The classification rule that underpins the scorecard outputs an estimate of the probability of belonging to class 1 at $\mathbf{x}$, $\hat{p}(1|\mathbf{x})$, with $\mathbf{x}$ feature vector. Standard practice involves modelling a specific definition with fixed definition threshold, $y=t$. Before the classification rule is constructed y is fixed so there is no scope for a change in definition once the classification rule has been constructed. If the credit practitioner wished to change the definition upon which their scorecard was based, a new model would be required.

Given that each account will correspond to some 'level' of creditworthiness we consider the cumulative distribution function of y, $F(y|\mathbf{x};\beta)$, with β a parameter vector. We propose a class of models which would permit us to change the definition after the initial classification rule had been constructed without requiring model re-estimation. This is possible by building the scorecard on a model based on the cumulative distribution of y given $\mathbf{x}$, $F(y|\mathbf{x};\beta)$. At the start of the model construction phase we do not fix y, as this would effectively reduce creditworthiness to a binary variable. Instead, the entire cumulative distribution is modelled. A model of this type is capable of producing probability estimates of $\mathbf{x}$ belonging to each possible value of y, so that for any choice of definition threshold the predicted probability of bad for that choice of definition may be calculated by integrating the probability density function. The predicted probabilities of belonging to the class of bad accounts are denoted by $\hat{p}(1|\mathbf{x})$. More generally the notation $\hat{p}_z(1|\mathbf{x})$ may be used to represent the predicted probabilities of belonging to the class of bad accounts when the definition threshold $t=z$ is used.

$$\hat{p}_t(1|\mathbf{x}) = \int_{-\infty}^{t} f(y|\mathbf{x};\beta)dx \qquad [1]$$

When a change in definition is required the underlying continuum may be partitioned by any value t of definition threshold. Using [1], the predicted probability may be calculated. Many models may be used in this manner. Perhaps the most appropriate candidates are generalised linear models (McCullagh and Nelder, 1983). Logistic regression, a special case of generalised linear models, is widely used in the credit industry and therefore may be particularly attractive for the credit scoring community. However, generalised linear models do not handle censored values as readily as survival and nearest neighbour methods as described below.

5.3.1 Survival Analysis

Survival analysis (described in Section 4.2) provides a suitable class of models for the implementation of global models. Traditional survival analysis compares the distributions of y for different feature vectors. We aim to use the model to predict the probability that an applicant with feature vector $\mathbf{x}$ is below some level t of definition threshold.

In credit scoring data are often censored. As discussed in Section 2.3.4.4, censoring occurs when the endpoint of an observation is not observed. For example, if the observed period of data is less than the loan term for unsecured personal loans, then any loan which is still active at the end of the observed data is censored. It can not be ascertained whether the loan will go on to complete repayments of the credit agreement or whether the loan would fall into arrears and be written-off as bad debt. Censoring can easily be incorporated into the survival analysis framework.

In Section 4.2 proportional hazards models were introduced and used as an alternative modelling tool. Here we can also use the proportional hazards models to calculate a global model because the cumulative distribution is modelled.

5.3.2 Nearest Neighbour Methods

Since we are solely interested in constructing a classification rule we can use nonparametric methods which avoid the model building stage entirely. For example nearest neighbour methods may be used, see Hand (1997), Devijver and Kittler (1982). Denoting the initial design sample by $(\boldsymbol{x}_i, y_i)$, $i=1,\ldots,n$, we find the k nearest neighbours to the feature vector of the new point to be classified in $\mathbf{x}$ space. Probability estimates of class membership may be obtained by calculating the proportions of these k nearest neighbours that have y values less than and greater than the definition threshold t. The value of t does not have to be specified until the classifications are required. Nearest neighbour methods have the added attraction that, if all censoring occurs on one side of any definition threshold

chosen, then the censoring causes no problems and may be ignored in the analysis. This situation arises with the UPL data sets discussed in Chapter 2. For the UPL(1) data set, if 3 years of data are available on 5 year loans, all loans still active at the end of the observed data period are censored at 3 years. Any definition threshold less than three years considers accounts still active as having not defaulted.

In using nearest neighbour methods, a metric (that is a distance measure used to assess the 'closeness' of observations) and k, the number of neighbours to base the analysis upon, must be chosen. The choice of k is to compromise between bias and variance (Friedman, 1997). If k is too big, large bias and small variance will result. k too small gives reduced bias but larger variance. In credit data we find that classification performance is relatively insensitive to choice of k (see Henley and Hand, 1997). Figure 5.1 shows the relative performance, in terms of error rate, for different values of k, with a loess smoothed curve plotted to aid interpretation. Many refined metrics have been developed, see for example Fukunaga and Flick (1984) or Myles and Hand (1990). These detail metrics that may be implemented with continuous data. Here we use the standard Euclidean distance metric, after a suitable scaling of the variables.

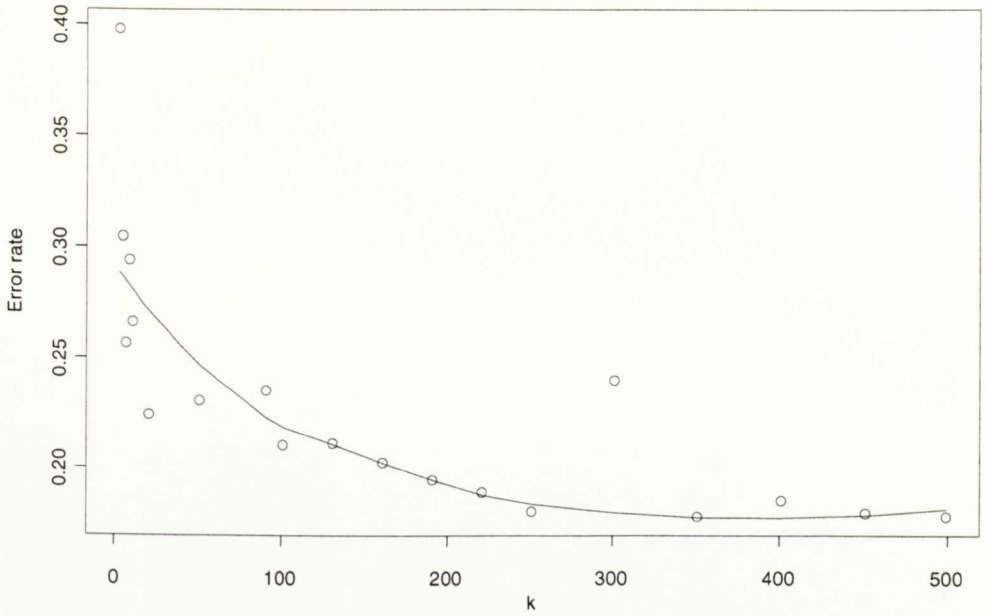

Figure 5.1: *Sensitivity of error rate with respect to choice of k*

Although choice of k is not critical some care must be taken. The choice of k should be made with reference to the class priors. If k is greater than $2m$, where m is the number of observations in the smaller class, then all observations would necessarily be assigned to the larger class.

5.4 Examples

Consider a purpose built classifier designed to suit one particular definition threshold. Perhaps the scorecard derived from this classifier has been used in the marketplace for a sufficient amount of time that the definition of a good account used in the scorecard may no longer be the ideal. Following the conventional approach remodelling may have been necessary.

Graphical illustrations of the predicted probabilities of belonging to the class of bad accounts derived from classifiers built using different definition thresholds highlight the possible differences the data may exhibit.

Three distinct possibilities are investigated. First, the situation where a new definition may be preferred but no classifier deterioration would be evident if the existing scorecard were used with data subjected to the new definition. Second, where apparent deterioration in classifier performance can be reversed by a change in classification threshold. Third, the situation where a global model would be the only alternative, to a complete classifier rebuild to prevent severe performance degradation.

5.4.1 No Action Necessary

The first example consists of observations extracted from UPL(1), described in Chapter 2. In light of the comments in Chapter 4 regarding the behaviour of different loan terms, here only accounts of term 24 months were used. In addition any accounts which settled ahead of schedule were excluded. This was because when using survival routines, such observations would be treated as censored. However, as noted in Section 4.2 these observations are an alternate endpoint whose characteristics differ slightly from those of censored observations.

The resulting data set comprises 3861 unsecured personal loan accounts of loan term 24 months. These data were divided into design and test sets consisting of 2685 and 1176 records respectively. Fourteen variables were used to formulate the prediction rule.

A definition for the prediction of profitability was proposed in Section 4.4. Here we consider different definition thresholds which may be implemented in the design of a scorecard. Under normal circumstances a bank may regard an application for an unsecured personal loan account as worth accepting if not more than 70% of the total interest payable is still outstanding at the end of the loan period. However, if competition for business amongst banks was high and the bank decided to try and increase its market share then a higher proportion of outstanding interest may be tolerated, say 80%. This example is for illustration – the definitions used are not intended to be of commercial significance.

Figure 5.2 shows how the distributions of the predicted probabilities from the two models differ. The vertical lines in each panel of the figure show classification thresholds, c, used to assign new observations to classes. The distribution of the predicted probabilities for observations $\hat{p}(1 \mid \mathbf{x}) < c$ clearly changes between the definitions. In Figure 5.2, c=0.3, in this case a change in the proportions of the population falling into either class is not evident.

Figure 5.3 illustrates a plot of default predicted probabilities assigned from the logistic regression model built with definition threshold 70% against the probabilities from the logistic regression model built with 80% as definition threshold.

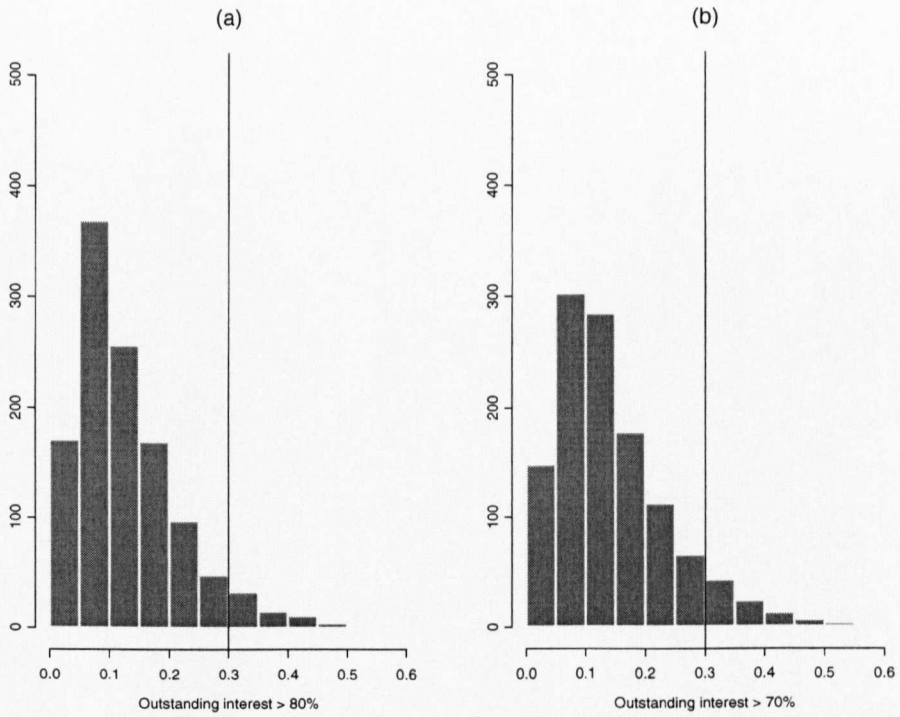

Figure 5.2: *Histograms of* $\hat{p}(1 \mid \mathbf{x})$ *for alternative definitions*

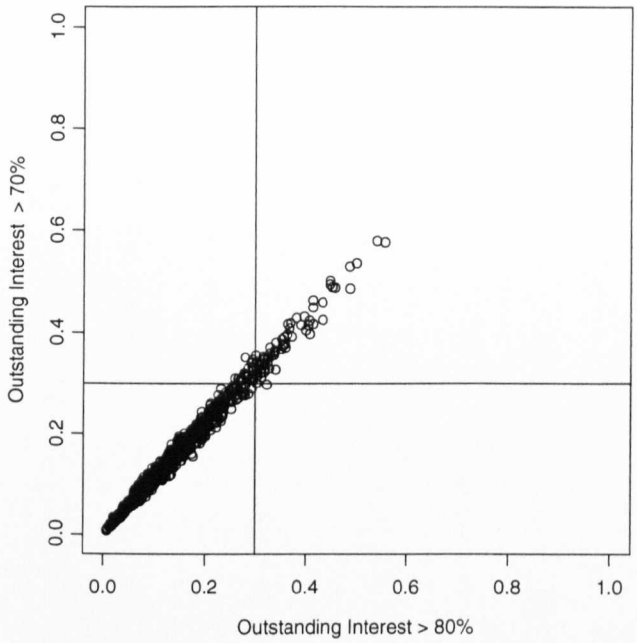

Figure 5.3: $\hat{p}_{70\%}(1 \mid \mathbf{x})$ *for model based on 70% against* $\hat{p}_{80\%}(1 \mid \mathbf{x})$ *for model based on 80%*

The simplest approach to deal with a change such as this in the definition threshold would be to ignore it completely. The classifier built for the 70% threshold would then require no alteration and would be reapplied to the modified data. The lines in Figure 5.3 indicate how the observations are classified when a classification threshold of 0.3 is implemented. It is clear from the diagram that most applicants are still assigned to the same class. The class assigned to observations in the bottom left and top right quadrants does not alter whichever definition threshold is used. The observations in the other two quadrants are classified into opposite classes according to the definition threshold used. The way in which these data behave under these circumstances leads to little deterioration of the classifier's performance. This is because the estimated probabilities from one model are approximately equal to the estimated probabilities from the other model. That is, $\hat{p}_{70\%}(1 \mid \mathbf{x}) \approx \hat{p}_{80\%}(1 \mid \mathbf{x})$. As a consequence very few points fall into the quadrants where the assigned class would be altered with changing class definition. The predicted probability of an

observation exceeding t=70% should be greater than that of exceeding t=80%. Note that this is demonstrated in Figure 5.3 by the observations being slightly above the diagonal.

Although no global model was required in this example, other applications of global models may prove useful. For example, a series of predictions could be obtained from a global model which suggest the subgroup of accounts most likely to fall in arrears to a certain degree yet extremely unlikely to default and cease paying altogether. Extra interest and penalties incurred on the arrears may result in this group of moderate defaulters proving to be the most profitable accounts. Traditional methods may have rejected this group of profitable accounts due to the possibility of missed payments.

It is likely that the classification performance resulting from a change in definition threshold is highly influenced by the magnitude of the change. To illustrate we replicate Figure 5.3 for ten possible definitions, t=[0.1,1] by 0.1.

Figure 5.4 clearly shows that the probability estimates arising from definitions 4 to 10 are linearly related. Any change in threshold between these definitions is not likely to result in significant change in classification performance. Should we wish to change definition where the probability estimates are not linearly related we must take a different course of action as outlined in the following sections.

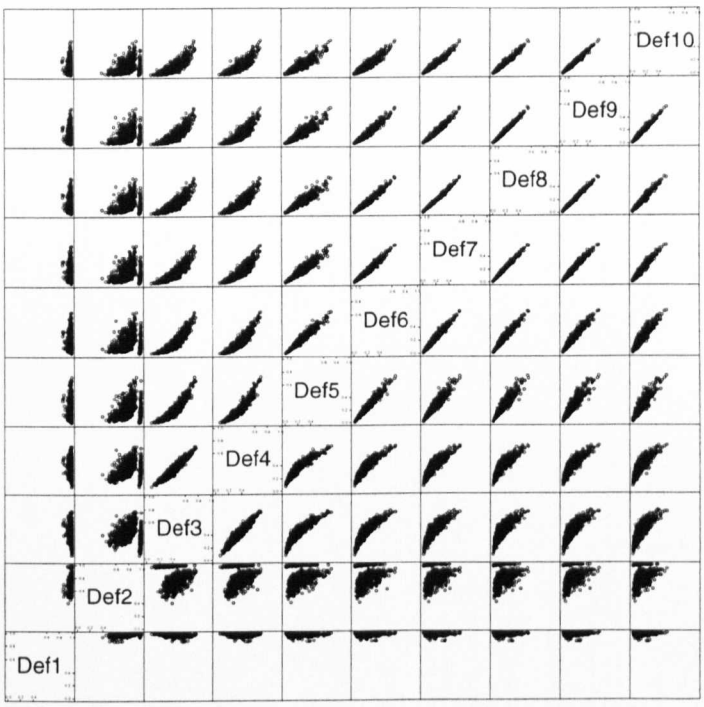

Figure 5.4: *Scatter plot matrix for* $\hat{p}_t(1|\mathbf{x})$ *arising from the UPL profitability data.*

5.4.2 A Change in Classification Threshold

This data set is derived from the current account data set described in Chapter 2. These data were randomly split into design and test sets of 3981 and 3975 observations respectively. The detailed description in Chapter 2 relates to practises similar to that which the data may be commercially used for. However, for illustrative purposes the framework in this chapter has been somewhat contrived.

Financial institutions often use complex definitions of good and bad current accounts, as described in Chapter 6. Here we adopt the simple definition that a desirable current account is one with a high debit turnover. In the extreme, if the debit turnover exceeds the credit turnover, (i.e. the account is in overdraft), then the bank can charge high levels of interest. Debit turnover is a continuous

variable and so can be partitioned at any level of definition threshold the bank chooses to define a good account.

The observations comprise monthly measurements for the following variables.

- average balance for the month
- credit turnover during the month
- balance at the end of the month
- minimum balance attained during the month
- maximum balance attained during the month
- current overdraft excess

In this example we utilise the data collected during months 0 to 3 to model the debit turnover for the sixth month.

In general the data may not have such convenient properties that the classifier requires no modification as in Section 5.4.1. If there is a large difference between the definition thresholds then the predicted probabilities of class membership are more likely to be dissimilar. If this is the case then the approach used in Section 5.4.1, that of continued use of the scorecard constructed using the original definition, is no longer useful.

We choose two definition thresholds that are widely spread in the definition threshold space. The definition thresholds used are £600 and £1600, where an account with debit turnover (DTO) less than the threshold indicates an undesirable account. Figure 5.5 illustrates the change in distribution of predicted probabilities between the two definitions used in this example. Imposing classification threshold c=0.5 in this case leads to many observations being assigned to a different class.

The predicted probabilities obtained using the definition thresholds as above are plotted against one another in Figure 5.6.

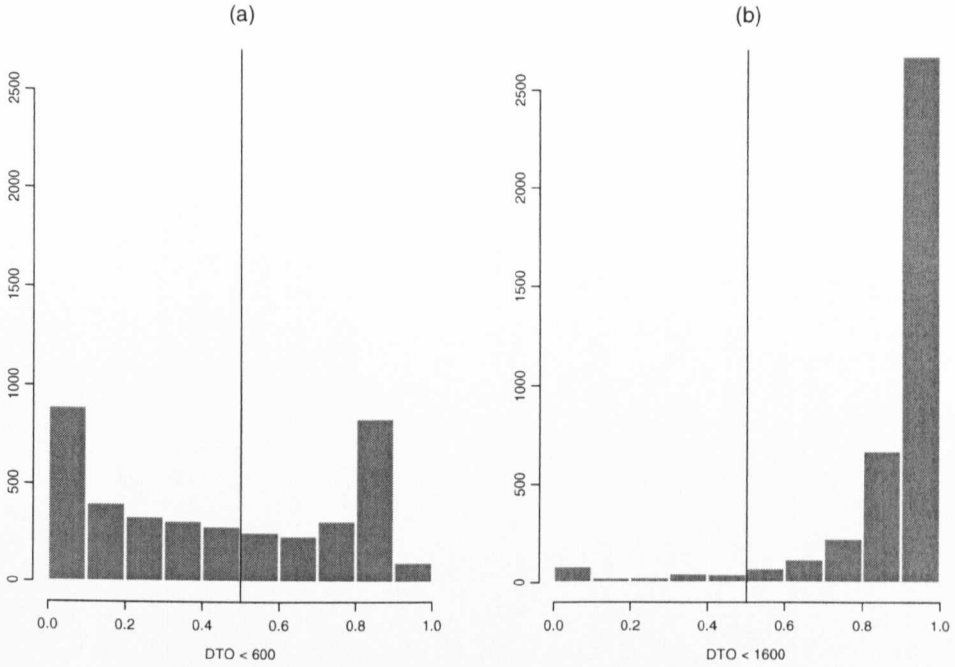

Figure 5.5: *Histograms of* $\hat{p}(1 \mid \mathbf{x})$ *for alternative current account definitions*

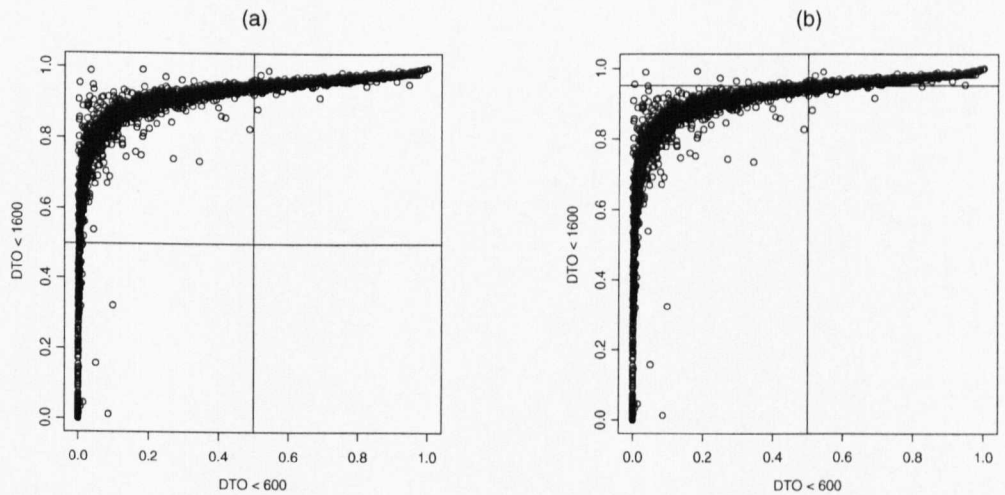

Figure 5.6: $\hat{p}_{£1600}(1 \mid \mathbf{x})$ *for model based on £1600 against* $\hat{p}_{£600}(1 \mid \mathbf{x})$ *for model based on £600.*

Figure 5.6(a) shows that if a classifier is constructed according to one definition when a second definition is believed to be more suitable then many observations would be incorrectly classified if the classification threshold remained unaltered

at 0.5. Using the definition threshold £1600 all the observations in the top left quadrant were assigned to the opposite class compared to when £600 was used. If no action was taken in light of a change in definition (as was the case with the data set examined in Section 5.4.1) results from each classifier would be very different. With such dramatic changes in class allocation one may expect extensive remodelling to be a necessity. However, the data exhibit behaviour which renders remodelling unnecessary – a change of classification threshold will suffice.

Figure 5.6 shows that the probabilities produced by the £1600 threshold can be approximated by a function of the probabilities generated when using the £600 threshold. For this to be true there must exist a monotonic relationship which can be applied to the probabilities generated by the first classifier in order to obtain the corresponding probabilities that would be obtained from the second classifier (if that classifier was constructed), i.e. $\hat{p}_{£1600}(1 \mid \mathbf{x}) \approx g\{\hat{p}_{£600}(1 \mid \mathbf{x})\}$, with g monotonic. The existence of such a monotonic relationship enables the same results to be obtained from each classifier simply by varying the classification threshold, rather than remodelling entirely. Figure 5.6(b) demonstrates that the results obtained by taking a classification threshold of 0.5 with definition threshold of £600 can be emulated by using a classification threshold of 0.95 with definition threshold £1600. If the probabilities could be exactly represented by a monotonic relationship, then varying the classification threshold of the second classification rule could reproduce the results exactly.

Figure 5.4 illustrated some of the possible relationships between the probability estimates when changing definition threshold. Figure 5.7 similarly shows ten combinations arising from different definitions when applied to the current account data.

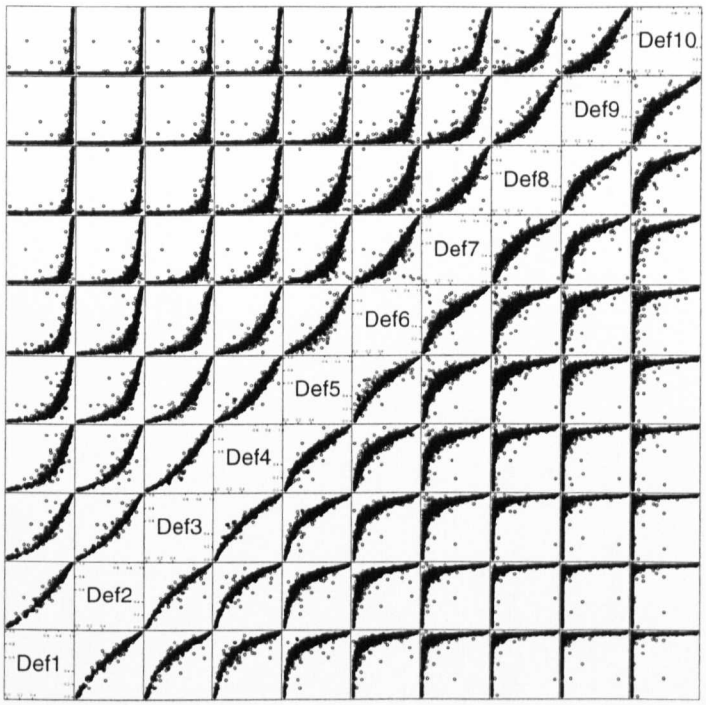

Figure 5.7: *Scatter plot matrix for* $\hat{p}_t(1|\mathbf{x})$ *arising from the current account data.*

Figure 5.7 clearly shows that there is a monotonic relationship between the probability estimates arising from most definition pairs. Those definitions not exhibiting a monotonic relationship are, in some sense, far apart. In Section 5.4.3 we provide examples to illustrate this situation.

5.4.3 Global Model Essential

Financial companies are often concerned that competitors may gain insight into their methods of scoring so that they may lose market advantage. In the credit data sets we have available, deterioration in classifier performance is only slight when the class definitions are changed within reasonable limits. However these methods should not be dismissed. Small differences in financial applications can result in substantial gains in profit. We illustrate two situations where global models would benefit the practitioner greatly. First, a simulated data set,

analogous to credit scoring problems, illustrates the effects of a global model. Second, a real data example where a global model is highly beneficial.

5.4.3.1 Simulated Data

The data were generated from a three component bivariate normal mixture, with components having mean vectors (0,0), (3,0) and (3,3), unit variance and zero correlation. Any observation, (x_1, x_2) from the first component had the value y=1 assigned, observations from the second, y=2, and observations from the third, y=3, where y is the underlying continuum which will be used to define the classes using definition threshold t. If the definition threshold is taken to be t_1=1.5, then the data for which y=1 can be regarded as good risks and observations for which y=2 or 3 as bad risks. When definition threshold t_1=2.5 is used, observations for which y=1 or 2 are assigned good risks with observations with y=3, bad risks.

A second simulation was performed using the above framework, with the three components from the bivariate normal mixture having mean vectors (0,0), (1,0) and (1,1). Figures 5.8(a) and (b) illustrate the structure of the data for the two simulations respectively.

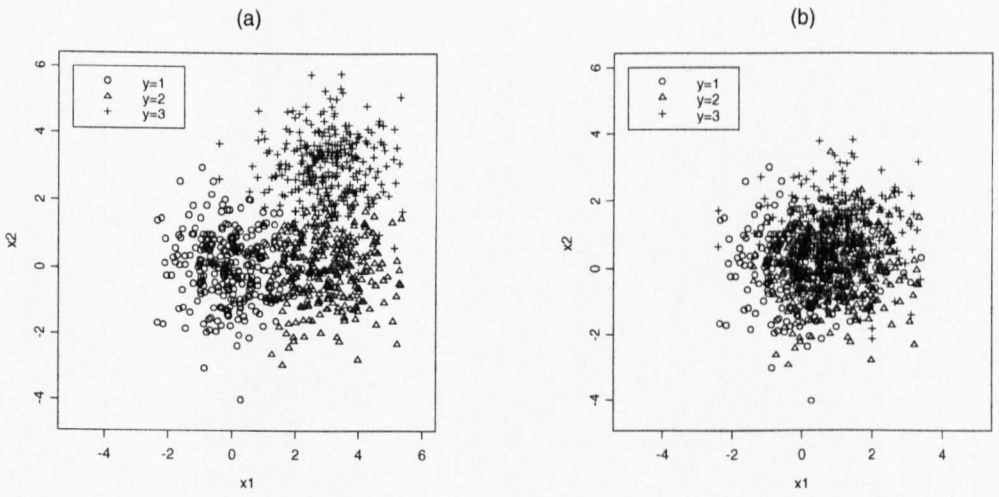

Figure 5.8: *Illustration of the spread of the simulated data sets.*

Utilising two definition thresholds and using the predictor variables x_1 and x_2 we construct separate logistic regression models for each value of t used to partition

these data. The logistic models are compared to a single global model constructed at the outset, in this case a nearest neighbour method. Logistic regression models were built using definition thresholds of t_1=1.5 and t_2=2.5. Figure 5.9 illustrates the relationship between the probabilities generated by these two different models.

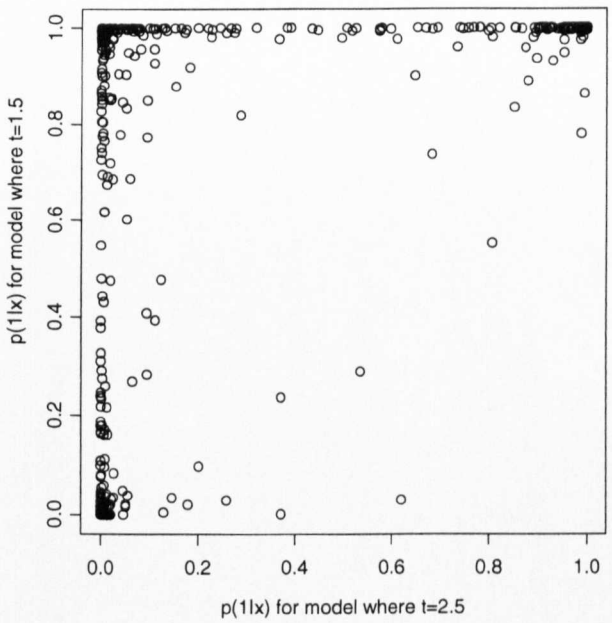

Figure 5.9: $\hat{p}_{1.5}(1 \mid \mathbf{x})$ *for model based on* t_1*=1.5 against* $\hat{p}_{2.5}(1 \mid \mathbf{x})$ *for model based on* t_2*=2.5*

Comparing Figure 5.9 with Figures 5.3 and 5.6 it is immediately apparent that this plot differs from the previous examples. Leaving the classifier unchanged as in Example 5.4.1 would lead to vast changes in class predictions. In Example 5.4.2 there was an alternative classification threshold that could be used to produce similar results when the second definition threshold was applied. These methods fail in this situation because there is no monotonic relationship between the two sets of predicted probabilities, $\hat{p}_{1.5}(1 \mid \mathrm{x})$ and $\hat{p}_{2.5}(1 \mid \mathrm{x})$.

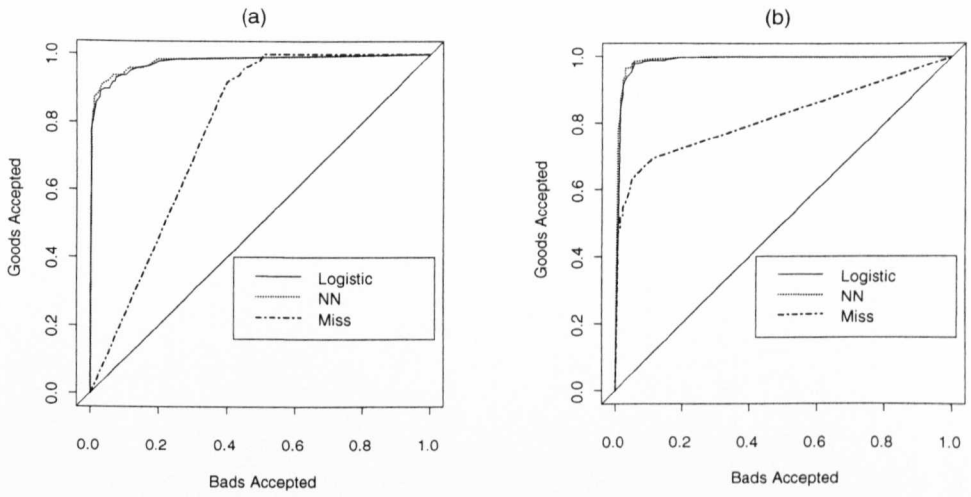

Figure 5.10: *ROC curves representing logistic regression, nearest neighbour, and misspecified logistic regression model for two definition thresholds.*

As discussed in Chapter 1, financial institutions commonly use ROC curves as an assessment mechanism for comparing the predictive power of different scorecards. There are several variants of ROC curves but here we shall use a plot of the proportion of good accounts accepted against the proportion of bad accounts accepted.

Panels (a) and (b) of Figure 5.10 respectively show three ROC curves relating to the two definition thresholds t_1=1.5 and t_2=2.5. The solid lines indicate the logistic models and the broken lines the nearest neighbour global models. The dot-dash lines are obtained by using the model built using t_1=1.5 as the definition threshold applied to the data defined using t_2=2.5 as threshold (Figure 5.10(a)) and vice versa (Figure 5.10(b)).

Both panels in Figure 5.10 clearly indicate that our global model, designed in a single step performs at least as well as both the logistic models. Furthermore they demonstrate that the classification performance will degrade severely should the methods outlined in Sections 5.4.1 or 5.4.2 be adopted.

While the classifiers built using these simulated data perform exceptionally well, classifiers in credit scoring rarely classify with this degree of accuracy. In order

to demonstrate that the arguments presented are equally valid with data which is not so readily separated into the correct classes we performed another simulation as detailed in Section 5.4.3.1 This time the three components from the bivariate normal mixture have means (0,0), (1,0) and (1,1). The corresponding plots to Figures 5.9 and 5.10 were obtained using this data set and displayed in Figures 5.11 and 5.12.

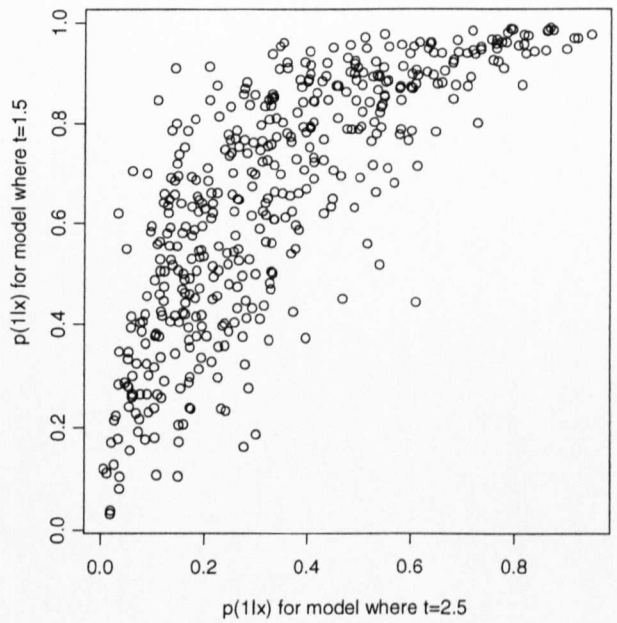

Figure 5.11: $\hat{p}_{1.5}(1 \mid \mathbf{x})$ *for model based on* t_1*=1.5 against* $\hat{p}_{2.5}(1 \mid \mathbf{x})$ *for model based on* t_2*=2.5 (second simulated data set).*

Panels (a) and (b) of Figure 5.12 respectively show three ROC curves relating to the two definition thresholds t_1=1.5 and t_2=2.5. In this case the data are not as separable yet the figures still clearly indicate that using a global model would be beneficial. Traditional techniques would again lead to costly classifier deterioration.

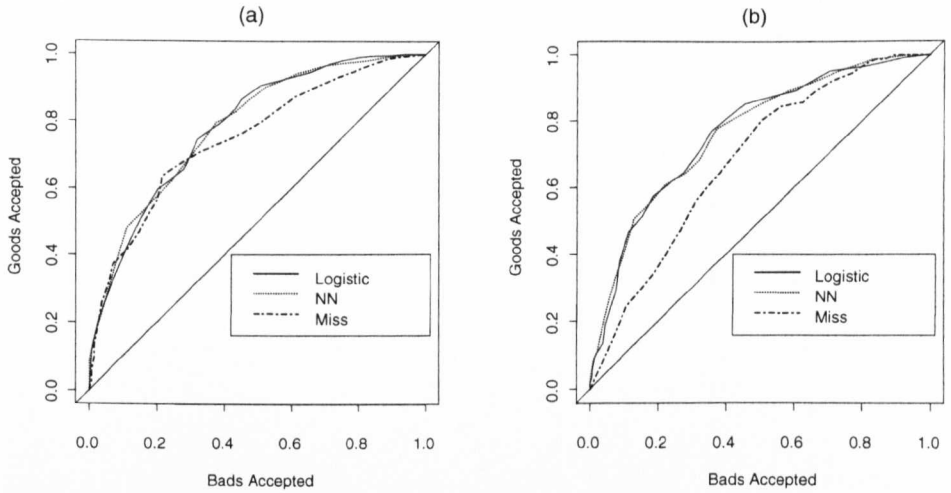

Figure 5.12: *ROC curves representing logistic regression , nearest neighbour and misspecified logistic regression model for two definition thresholds investigated (second simulated data set)*

5.4.3.2 Real Data Example

As we have seen in this chapter, monotonic relationships exist between the sets of predicted probabilities for the available data. Consequently, a global model is not essential for our available credit data sets. That is not to say that other credit examples may exist where global models prove hugely beneficial. To illustrate a situation where global models do prove useful we use the following example from engineering.

The data set is a sub-sample extracted from a larger data set provided by the Defence Research Agency. The data comprise measurements of energy measured at two polarisation angles for two different wavelengths of light shone on a compound of silicon and germanium. The response variable y is the proportion of germanium in the top 10 angstroms of the specimen, and can take values from 0 to 0.15. Each observation can be assigned to one of two classes. The class is dependent on whether the proportion of germanium exceeds some threshold - a definition threshold. The classification rules were formulated and assessed using design and test samples, each consisting of 2000 observations. Figure 5.13 shows a plot of the predicted probabilities from each definition.

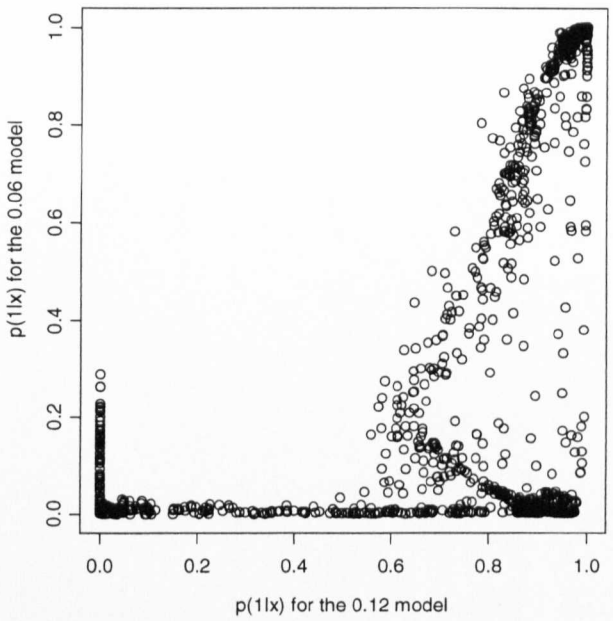

Figure 5.13: $\hat{p}_{0.06}(1 \mid \mathbf{x})$ *for the model based on* t_1*=0.06 against* $\hat{p}_{0.12}(1 \mid \mathbf{x})$ *for the model based on* t_2*=0.12 for the second simulated data set.*

From Figure 5.13 it is clear that there is no monotonic relationship between the predicted probabilities generated for the two choices of definition threshold. Regardless of the choice of classification threshold severe performance deterioration will result if the definition threshold is changed from 0.12 to 0.06. Figure 5.14 show the results obtained from ROC curve analysis when the two definition thresholds are used.

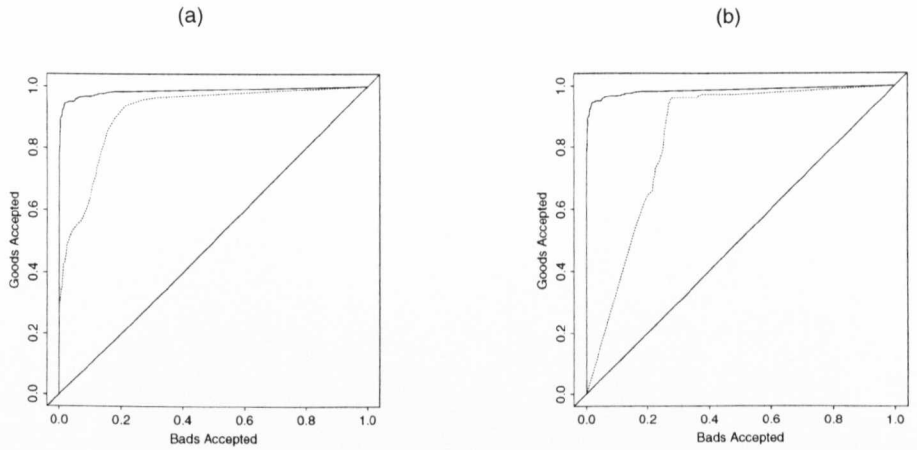

Figure 5.14: *Nearest neighbour global model classifier compared with misspecified logistic regression classifiers, (a) t= 0.12, (b) t=0.06).*

The global model used for this analysis was a nearest neighbour model with k=101 and Euclidean distance metric. Initially models were constructed to compare classifiers based on nearest neighbour analysis and logistic regression. ROC performance was very similar between classifiers. However, Figure 5.14 shows that if a logistic classifier constructed using either definition continued to be used after a change in definition then classification performance would worsen. On the other hand a single global model would maintain performance at a high level even after a change in definition threshold.

5.5 Conclusions

In this chapter we have discussed the ideas and methodology and implemented several examples of global models that can be used in credit scoring applications where the class definitions are subject to a degree of flexibility. We considered:

1. Whether the classification performance of a scorecard degrades if the class definitions are altered?

2. If (1) is true, does a *global model* provide better results?

3. If (2) is true, how does the performance of the global model compare to a classification rule built specifically for the new definition?

From our results it is reasonable to suggest:

In some circumstances the classification performance of a scorecard can deteriorate in a manner which cannot be compensated for through adjustment of the existing classifier. When this is the case we have seen that global models provide a suitable alternative that performs at least as well as standard purpose built models for each bad definition in our example.

A global model allows the user to build a single model at the outset and vary the class definitions when the classifications are required. Standard procedure would require reconstruction of the classification rule according to the new definition.

Chapter 6

Definitions and Optimal Models

6.1 Introduction

In Chapter 5 models which allow flexibility when deciding how to classify new applicants were introduced and discussed. The global models we describe are based on the notion of a single underlying continuum y, typically some representation of creditworthiness. This continuum is partitioned in order to produce two classes of interest. Flexibility is provided by modelling the cumulative distribution of y given $\mathbf{x}$, $F(y|\mathbf{x};\beta)$, where $\mathbf{x}$ is a feature vector. This allows the classes to be defined at the time classification is required.

In other areas of the credit domain class definitions are often more complex. Rather than being based on a single variable they can involve several variables. Moreover, each of the variables involved in the definition typically has a certain arbitrariness associated with them, as was the case with examples described in Chapter 5. Adopting the global approach enables models to be constructed which may be varied in order to produce different predictions of class membership according to the characteristics of the problem at the time of classification. When multiple variables contribute to the definition, the framework of the problem is much more complex and it is not easy to usefully extend the ideas of global models in such situations, for example a multiple response survival analysis model. Further research in this area is required to develop multivariate global models.

Li and Hand (1997), Hand, Li and Adams (1998) adopt a different approach to such problems – *indirect models*. Indirect prediction models are built to predict intermediate variables, which are the collection of variables underlying the definition. These may then be used in deterministic class definitions. One advantage of this approach is that in the event of a change in the class definitions, a completely new model need not be constructed.

Global models are an ideal solution when the formulation of a problem is such that a change in the definition of the classes to be predicted may be appropriate after a certain amount of time has passed. Examples throughout this thesis have involved unsecured personal loan accounts. A customer who misses a payment on a loan such that they are breaking their loan agreement is clearly not an ideal customer. The arbitrary nature of these problems arises when deciding how large the breach of contract must be in order for the customer to be sufficiently undesirable that the bank will seek to exclude such customers from their account base. With definitions involving more complicated combinations of variables we find that even those with banking expertise do not have a clear interpretation of what they actually regard as bad. Here, we may find that part of the bad definition incorporates how long an account has been dormant. An account which is not credited during a two month period certainly may be undesirable from a business perspective, however no agreement has necessarily been violated. Other similar judgements are incorporated into credit definitions used in industry and will be discussed in the example below.

Our example is typical of a current account definition, formulated using several variables. The precise definition formulation has been modified to preserve confidentiality. Each variable that contributes to the bad definition implemented by the bank is partitioned. A change in any one of these partitioning thresholds would give rise to an alternative definition. Moreover, the very fact that these thresholds are somewhat arbitrary (since no definitive definition exists), suggests that providing the thresholds are not varied too much, there is no reason to suspect that any alternative definition would be any less valid than the original.

Furthermore, a region of definitions is defined by fixing limits between which each variable can take its value. This region of possible definitions can be reduced to a computationally manageable number of definitions by varying each of the variable thresholds by fixed increments between the stated limits. In this chapter the performance of definitions lying in this region will be examined. In addition, criteria that may be used to determine a suitable choice of definition to be used in a commercial environment will be proposed.

6.2 Traditional Approach to the Problem

Here we utilise the current accounts data set, described in Chapter 2. The data consist of variables which are recorded monthly, on 7956 accounts. Current practice is to construct a model to predict whether an account will be classified into the bad class by a certain time point in the future using a definition such as Definition 1, given below.

Using the variables given below, a classification rule based on logistic regression may be constructed. This approach would be very similar to the methods used in a bank to model bad accounts.

Four variables are combined in the formulation of the bad definition:

- Max balance attained during month (MaxBal)
- Monthly end balance (EndBal)
- Current excess (Excess)
- Credit turnover (CTO)

Note that only the ratio of current excess and monthly end balance is required in Definition 1.

For commercial reasons the definition we use is not identical to that used in practice, but has been slightly modified to give Definition 1; as follows:

Definition 1:

An account is classed as bad if at least one of the following are true:

i) Excess > £500

ii) Excess > £100 and MaxBal< 0

iii) CTO/EndBal< 0.10

Excess in i) and ii) will, henceforth be referred to as Excess(i) and Excess(ii) respectively.

Five variables used to predict Definition 1:

- Debit turnover(x_1)
- Number of cheques(x_2)
- Number of direct debits(x_3)
- Value of debits(x_4)
- Value of charges(x_5)

The approach used by banks would typically be to construct a scorecard using logistic regression according to the above definition. Following this procedure produced a classification rule with a Gini coefficient of 0.409.

Such approaches are in common use throughout the credit industry. The results produced by such analysis are regarded as the norm by credit practitioners.

We have stated that the definition upon which this analysis is based certainly has fairly arbitrary thresholds for the variables. For example, the 'roundness' of the figures used in the formulation of Definition 1 leads one to suspect that at least part of the definition was chosen by a domain expert rather than based on any formal analysis. If the partitioning thresholds have been rounded then it would be reasonable to suggest that many other values for these thresholds would give rise

to alternative, equally valid class definitions. These alternative definitions may be equally acceptable to the bank. For example, bankers may not resist changes such as Excess(i) > £450 or Credit Turnover < 15% of Monthly end balance being incorporated into the definition instead of the values given in Definition 1. If this is the case then the consequences of such changes on classifier performance are potentially interesting.

6.3 Changing Definitions

Our approach is to explore definitions that lie within certain limits for each of the variables that feature in the definition. We limit the search space to a computationally manageable amount by constraining the variables to take values at regular increments between upper and lower bounds chosen by a domain expert. Incorporating prior knowledge in this way ensures that any definition found to be desirable in terms of classifier performance would also have a sensible business interpretation. Care should be taken to choose a search space wisely, since tens of thousands of definitions can easily be generated by choosing a wide definition search grid and small incremental increases.

Denoting the partitioning values of the definition variables Excess(i), Excess(ii), Max Bal, and the ratio of Credit Turnover and Monthly End Balance as t_1, t_2, t_3 and t_4 respectively. We examine classification rules constructed according to the definitions which arise when the variables are varied between the limits $t_1 \in [200,800]$, $t_2 \in [50,600]$, $t_3 \in [-150,150]$ and $t_4 \in [0.05,0.5]$.

Varying the variable thresholds by increments 100, 50, 50 and 0.05 for t_1, t_2, t_3 and t_4 respectively leads to 5880 possible distinct definitions. These increments were chosen in accordance with the results of a pilot study, and in a manner thought adequate for the definition region for this problem.

A separate logistic model was built for each of the 5880 definitions in the above search grid. Initially the models were estimated using a randomly selected half of

the data, and performance assessed with the remaining half. However, a subtle problem caused results from this analysis to be unreliable. The choice for the final definition was influenced by the test set. These matters are described in the following section. Thus, the method of model estimation required modification.

6.4 Definition Choice

Constructing and validating a classification rule using design and test sets each consisting of approximately 4000 observations with 10 predictor variables is not likely to produce an optimistic classification rule. However, this problem introduces a different kind of overfitting that is seldom encountered in classification problems. In this example numerous definitions are modelled, and assessed using an independent test set. Test set performance on the basis of Gini coefficient is then used in selecting the definitions that perform well. This extra flexibility may increase the possibility of overfitting. Effectively the performance for each classifier on a single independent test set is used as a further criterion for definition choice. This could bias the final choice of definition.

A suitable refinement in the estimation procedure would be to use *cross-validation,* Efron and Tibshirani (1993). Cross-validation involves splitting the data into n subsets. A classification rule is constructed using data from n-1 subsets and the remaining subset used to assess the performance of the classification rule. This process is repeated n times and performance results are averaged. Provided n is not too small, n-1 represents most of the data, so that each classification rule constructed will be similar to that constructed using the whole data set. We used ten-fold cross-validation. That is, split the data into ten subsets and take 9/10ths of the data to use as design set each time. This is repeated ten times and the average of the test set results used as the performance measure for each definition. This approach eliminates any overfitting problems because the criterion used to select between definitions is the average of the

performance of ten independent models validated using different test sets (the random tenths of the data sets).

Other approaches include, leave-one-out cross-validation which involves many model building repetitions using a single observation as test set, or using a bootstrap approach, which draws random samples from the data with replacement. The first of these alternatives would be too computationally demanding. The second, unlikely to provide a significantly better result due to our large initial sample sizes.

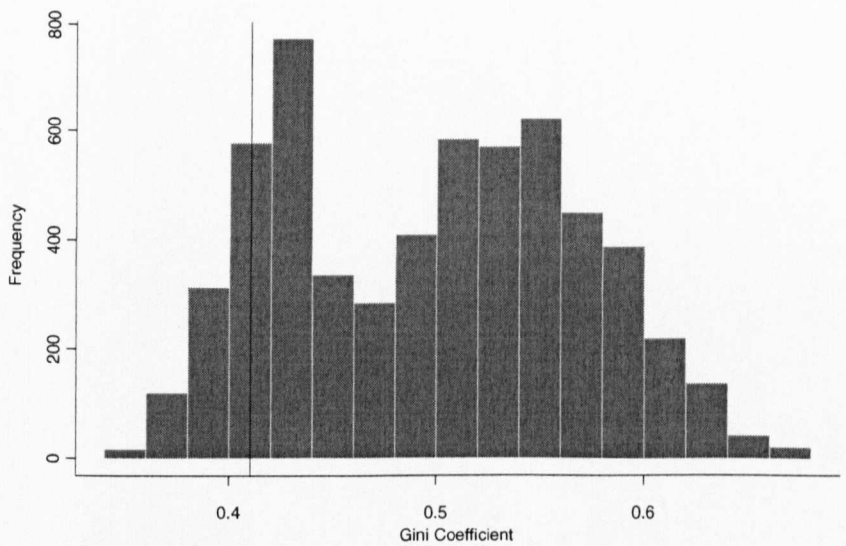

Figure 6.1: *Distribution of Gini Coefficients corresponding to the 5880 Definitions.*

The distribution of the Gini coefficients obtained from the logistic classification rules constructed according to the 5880 definitions is illustrated in Figure 6.1. Prior to this analysis, there was no reason to suspect that the Gini coefficients for this region of definitions would suggest a bimodal distribution as in Figure 6.1. Indeed, the apparent bimodality is an artefact of the search grid used to explore the region of definitions. Given that bankers may be satisfied to accept any of the 'equally acceptable' definitions investigated, a low standard deviation among

definitions may have been anticipated. The mean Gini coefficient is 0.498 with standard deviation 0.071.

The Gini coefficient corresponding to Definition 1, indicated in Figure 6.1, is situated towards the left tail of the distribution. Definition 1 produces a classification rule whose performance is much poorer than the majority of the alternative definitions. This being the case, definitions in the high valued tail are of interest.

Other than the actual Gini coefficient attained by the models constructed according to different definitions, the proportion of the population classified as bad by the definition may also be of interest. Classification rules for credit data can seemingly perform artificially well when assessed using a criterion such as error rate. Simply classifying all observations to the larger class is likely to provide a low error rate. However, classifiers which correctly predict bad accounts (with a higher overall error rate) are likely to provide scoring systems which are more financially viable.

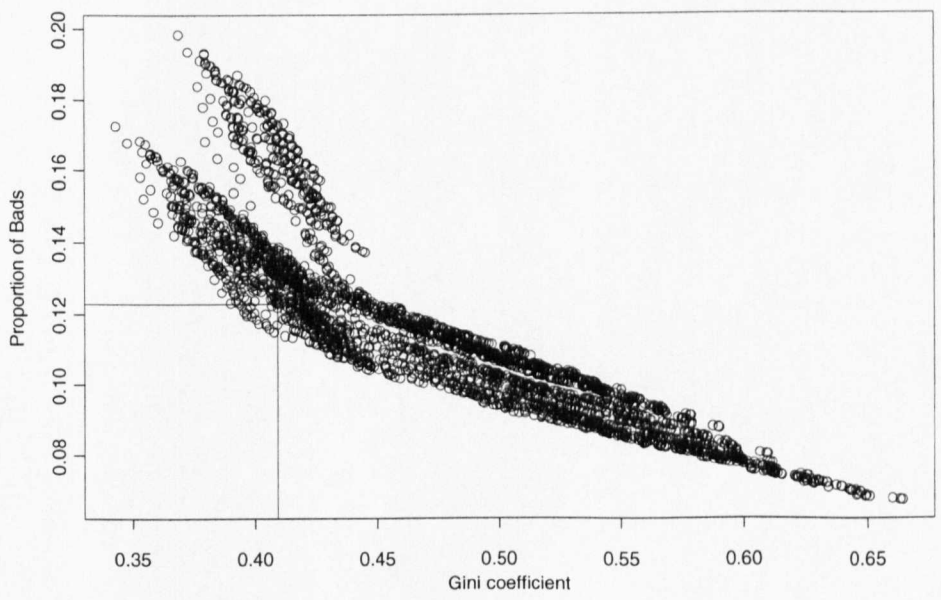

Figure 6.2: *Gini coefficients obtained from 10 fold cross-validation plotted against proportion of bads for the 5880 definitions*

Figure 6.2 shows that the Gini coefficients are distributed into two clusters. The solid black line indicates Definition 1.

6.4.1 Competing Definitions

Figures 6.2 shows that the performance of Definition 1 lies in the left tail of the distribution of Gini coefficients – a low value compared with the majority of the definitions considered. Many equally acceptable choices of definition would have produced classes that could be predicted with a higher degree of accuracy. Definition 1 has Gini coefficient equal to 0.409, the obtainable maximum in the 5880 definitions is 0.65. Table 6.1 contains information on Definition 1 and three alternatives.

	Definition 1	**Definition 2**	**Definition 3**	**Definition 4**
t_1	500	400	200	600
t_2	100	150	100	400
t_3	0	-50	150	0
t_4	10	5	10	5
Gini coefficient	0.409	0.464	0.360	0.612
Percent bads	12.290	10.210	14.560	7.470

Table 6.1: *Alternative Definitions*

The alternative definitions in Table 6.1 each have thresholds different from Definition 1. It would be very difficult to argue that Definition 2 is any less valid than Definition 1. Definition 2 yields a Gini coefficient of 0.464 – a marked increase from the original definition. It is by no means obvious which of Definitions 3 and 4 one might expect to lead to the largest Gini coefficient. Even bank experts did not have an intuitive feel for the differences in performance that emerged. Both definitions have common variable values with Definition 1, both have thresholds more extreme than Definition 1. Definition 3 gives a less favourable model with Gini coefficient equal to 0.360, whilst Definition 4 gives 0.612 Gini value – an enormous improvement.

Table 6.1 also indicates that Gini coefficient is inversely proportional to the size of the bad class, as noted by Hand and Kelly (1998).

All 5880 definitions are formulated using four variable values. Figure 6.3 and 6.4 illustrate how the values of these four variables effect Gini performance and class proportions, classified by the definition.

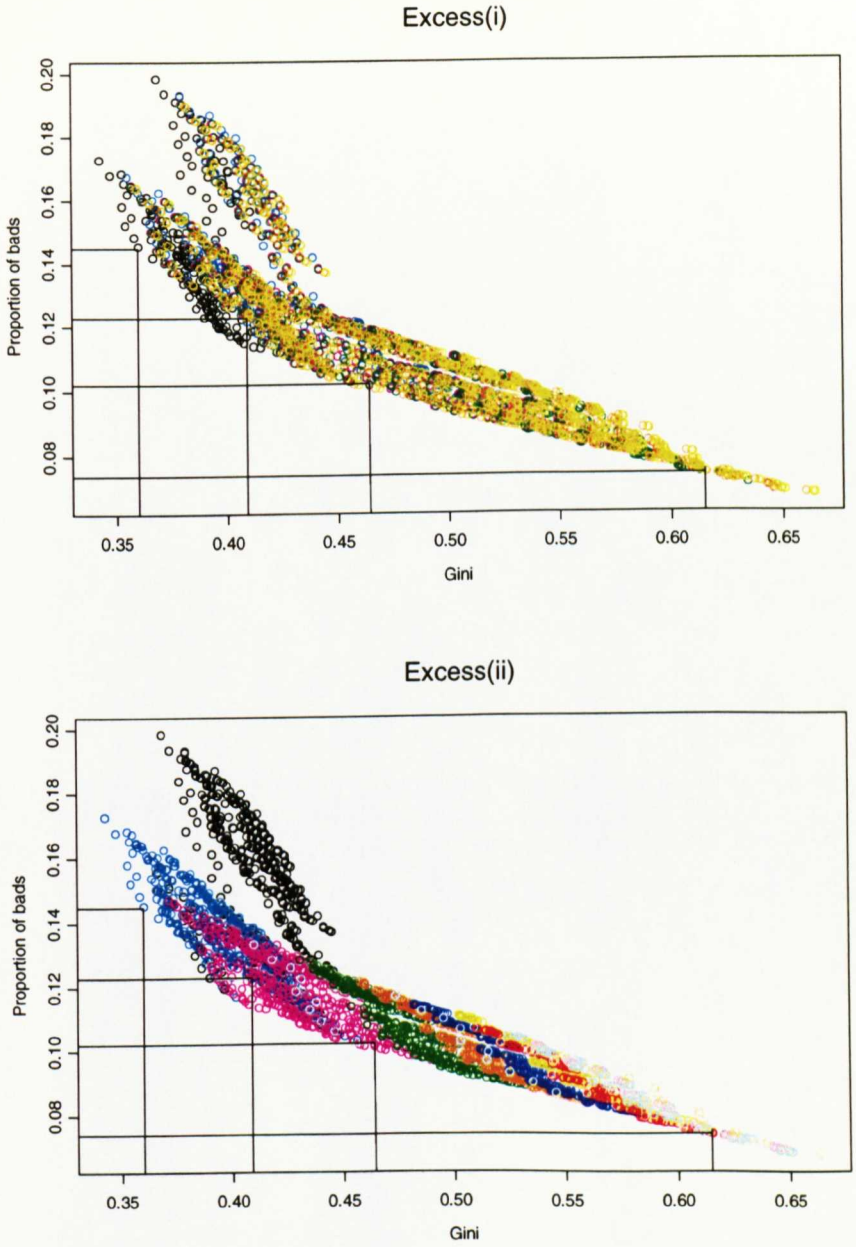

Figure 6.3: *The effect of definition variables Excess(i) and Excess(ii) on classification performance.*

Duplicates of Figure 6.2 are shown in Figures 6.3 and 6.4, each panel represents one of the variables which contributes towards the bad definition. Observations are coloured according to the value taken by the respective variable used in the definition.

As noted previously the spread of Gini coefficients forms two clusters, one much larger than the second. It can be seen from Figure 6.3(b) that the common feature of the definitions contained in the small cluster is that the majority have Excess(ii) equal to £50. This apparent bimodality may be explained by the constraints on the values which Excess(ii) took in the analysis. Examining Figure 6.3(b) which corresponds to Excess(ii), it is clear that the effect upon the Gini coefficient when altering Excess(ii) from 50 to 100 has a much larger impact than when Excess(ii) is altered by £50 in the rest of the search space.

Figure 6.3(b) clearly demonstrates that choice of Excess(ii) has a tremendous impact on the Gini coefficient of the resulting model. Where as Figure 6.3(a) fails to provide a coherent message of the effect varying Excess(i) has on Gini coefficient. Figure 6.5 provides a clearer visualisation for this variable.

The ratio of credit turnover to months end balance, Figure 6.4(b), also has a marked effect on separability. Many of the values taken by this variable encapsulate definitions whose range of Gini coefficient is large. Clearly as CTO/EndBal increases the range of Gini coefficient attainable from the corresponding definitions is decreasing. Note that this variable is particularly influential on the proportion of bads. No easily interpretable structure can be seen for MaxBal in Figure 6.4(a).

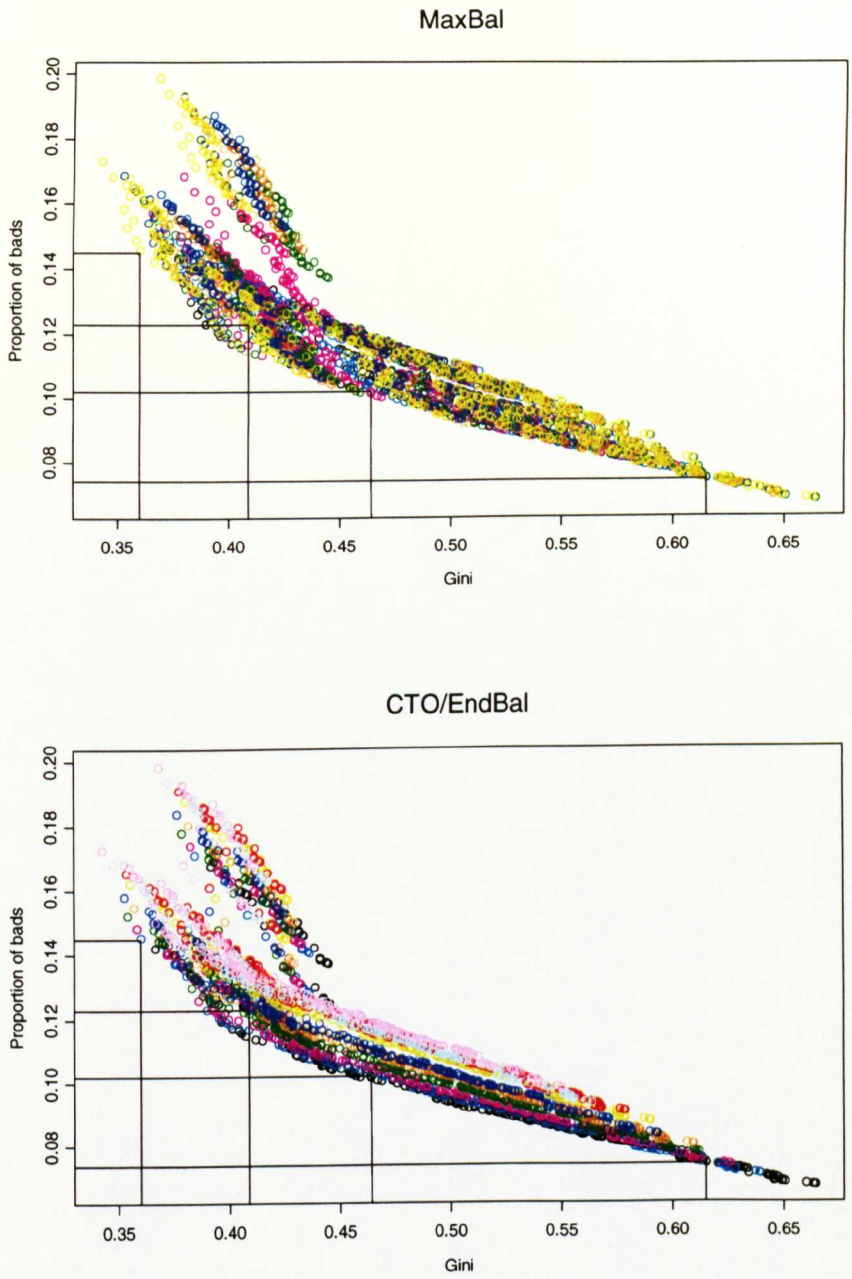

Figure 6.4: *The effect of definition variables MaxBal and CTO/EndBal on performance.*

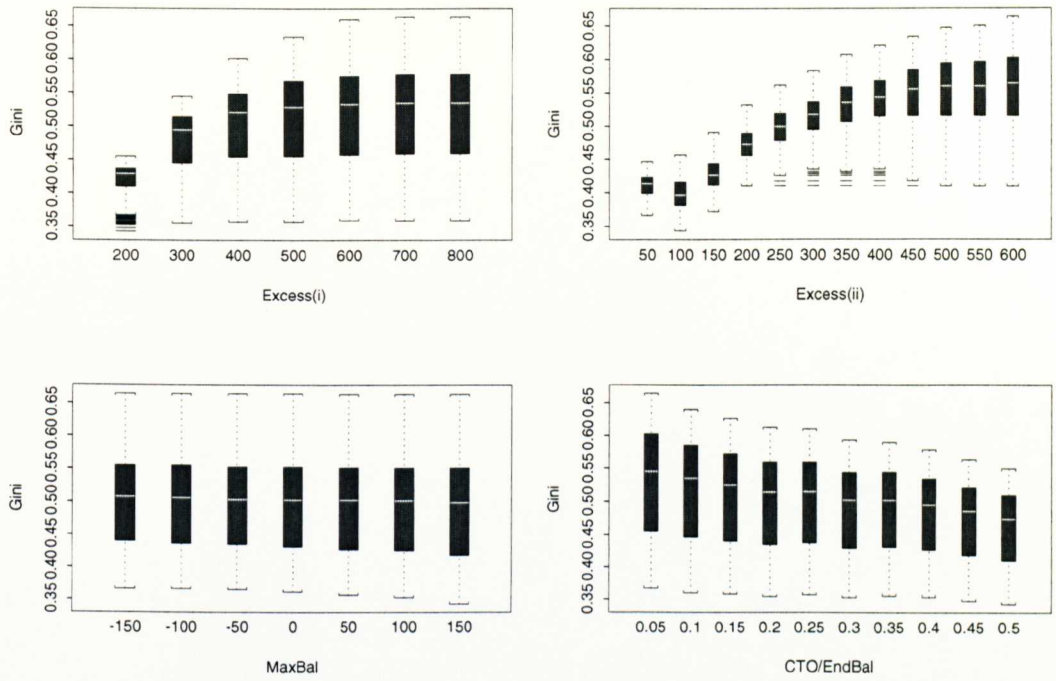

Figure 6.5: *Box plots of cross-validation Gini coefficient for each value of each variable used in the formulation of the definition.*

Figure 6.5 shows that low values of Excess(i) are related to low Gini values. This is not apparent from Figure 6.3 due to observations with different values being coincident. Also note that increasing Excess(i) has no effect on the Gini coefficient once the value reaches £400. The box plots for Maxbal reinforce the view that this variable has little impact, regardless of the value it takes (in the range investigated). It is clear that for Excess(i) and CTO/EndBal, certain values would, if included in the definition, prove advantageous in terms of the discriminatory power of the model.

Figure 6.5 also shows that the variability of the distribution of Gini coefficients can vary by a large amount as the variable values are changed. An increase in variability is particularly striking with increasing Excess(i) and Excess(ii).

At present the definition is quite simple, only one bivariate interaction is included in the definition of bad. However, there is nothing to stop higher order interactions being included. Inevitably this sort of practice would lead to even

more definitions that may be considered for the problem. If more variables were included in the definition, the previous search space would be a subset of this new search space incorporating more interactions. The resulting classifier performance from this new search space must necessarily be at least as good as previous studies. Note that these kind of changes require banking expertise to ensure that new variables introduced to the definition, and different combinations of variables, are sensible.

In conclusion, we have shown that the performance of the classifier built using Definition 1 is poor when compared to the alternative definitions in the surrounding neighbourhood.

6.5 Model performance

To build an operational scorecard requires a specific classes definition. Essentially there are two ways to view this problem. Firstly, appealing to the fact that there is no definitive definition, each of the 5880 definitions may be regarded as equally acceptable. This being the case, one can regard all definitions as in competition with one another. The definition which performed best on the chosen criterion would be chosen for commercial implementation. This course of action may be regarded as quite radical when compared with traditional banking procedures.

Bankers are often reluctant to implement such a radical change from their existing methods. A second, more conservative approach that may be preferred in such circumstances is to use their existing definition as a basis for comparison. Assessment of a new definition is gauged both by Gini coefficient (for example) and how closely related it is to the original definition. This approach assigns Definition 1 some degree of credibility and uses this definition as an idea about which to relate any definition which is perceived to be superior.

Note that in the analysis above, the standard definition was used, in some sense, as the centre of the search grid. However, the centre of the search gird was determined such that all surrounding definitions would have plausible business interpretations, but no special meaning was assigned to that definition

These two distinct approaches may be summarised.

The exact notion of a bad current account is unknown. The solution to the problem is obtained by selecting the best possible definition from a group of plausible definitions by maximising some external criterion, in our case Gini coefficient.

One particular definition is preferred and so a compromise is made between deviation from the preferred definition and improvement in classification performance. This compromise is equivalent to stating that it is more useful to predict the new definition with a greater accuracy than it is to continue using the old definition with current accuracy.

It has been shown that varying the definition can lead to vastly different values of Gini coefficient. If determining a definition on the basis of comparison with Definition 1, high Gini coefficient alone is not necessarily related to high definition desirability. A measure of closeness to the reference point is required.

A measure of this type may compare the classes to which the observations are allocated by each definition. The proportions of observations common to the good classes, or the bad classes, or total observations common to both classes give three possible measures which could be used to compare alternative definitions with the standard. The proportion of goods and bads common to the classes arising from different definitions will be referred to as the good/bad distance (GB distance).

These three distance measures are related to Gini coefficient as demonstrated in the figures below. The four points on Figures 6.6 and 6.7 marked by squares

centred on a large cross correspond to definitions 1 to 4 that have previously been described.

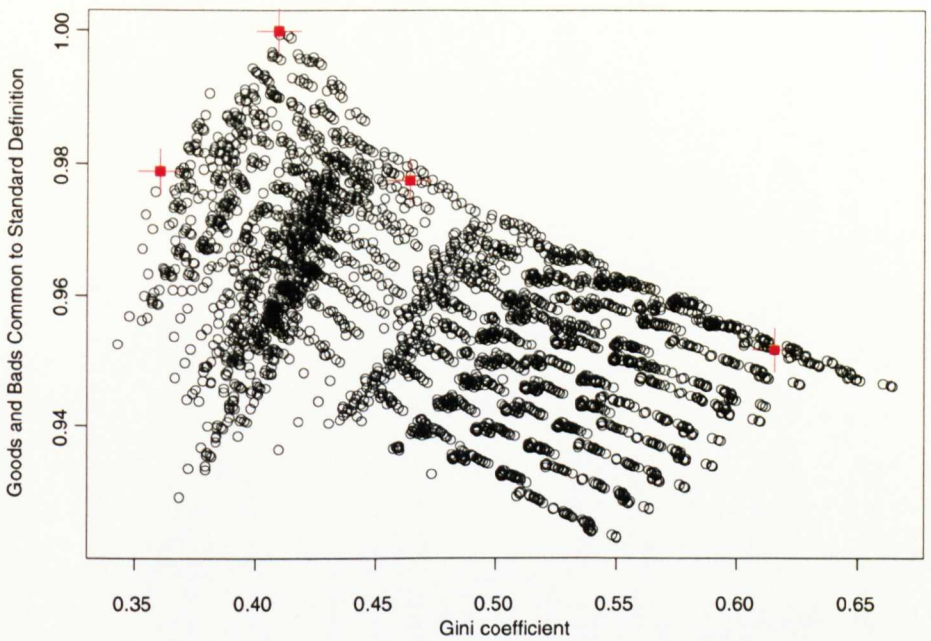

Figure 6.6: *Relation of Gini Coefficient and Class Commonalties for Alternative Definitions*

A reasonable measure of distance from the standard definition is the GB distance. This measure would allow statements such as, "using this alternative definition increases classification performance in terms of Gini coefficient by 0.25 whilst 97% of observations are assigned to the same class whichever definition is used."

However, using GB distance alone is artificially optimistic. In credit scoring the bad class is usually much smaller than the good class. The maximum and minimum proportion of bads in the definition space explored here are 19.85% and 6.75% respectively, with the standard definition resulting in 12.29%. GB distance may result in 97% of observations being assigned the same class under the new definition, but when examined more closely, 100% of the goods may be common yet only 65% of the bads. If the bad class is the most interesting and, perhaps more importantly the most costly when misclassified, then this drop in

the proportion of the population actually defined as bad may not be an acceptable compromise to make simply to gain more separable classes.

Figure 6.7(a) and (b) illustrate the analogous plots to Figure 6.6 using only the proportions of goods and bads respectively.

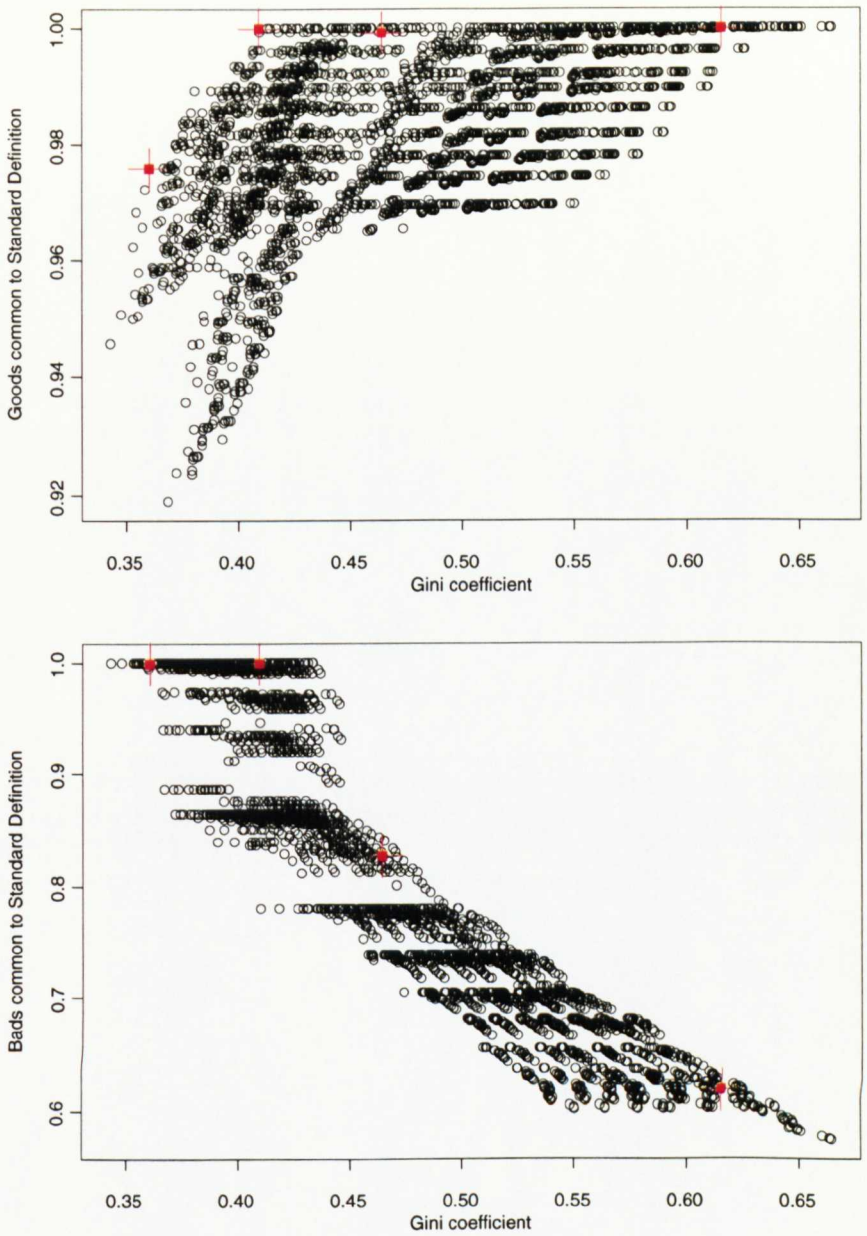

Figure 6.7: *Panels (a) and (b) Good and Bad Commonalties between definitions*

These plots can be used to make a decision for a new definition on more restricted grounds. The bank may wish to impose a constraint on the new definitions that they will consider. For example, they may consider only definitions where at least 85% of the bad class from the existing definition are also classed as bad by an alternative definition. Of course if the amount by which a definition can differ from the original is constrained then improvements in classification performance are not likely to be of the same magnitude as unconstrained definitions. However, even in this example we see a possible range of Gini coefficients between 0.371 and 0.448 – this would still be regarded a tremendous improvement in performance. Figure 6.8 incorporates all three distance measures.

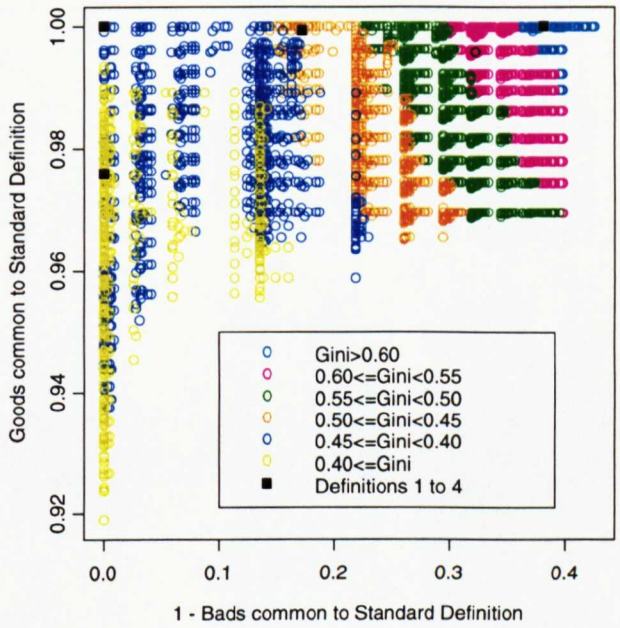

Figure 6.8: *Comparison of model performance*

Plotting common goods against 1- common bads gives Figure 6.8. Definitions closer to the top left hand corner of Figure 6.8 indicate large GB distance. This figure is coloured according to Gini coefficient. This enables a range of obtainable Gini values to be specified according to any limitations on definition

acceptability, specified by the bank. For example, if definitions were only acceptable to the bank if between 98% and 96.5% bads classed by the new definition were common to the original, Figure 6.8 may be used to state that Gini coefficients in the range (0.40,0.60] are obtainable subject to this aditional criterion.

6.5.1 Graphical Representation

When it is desirable to compare any proposed definition with the existing definition the following triangle plots may be useful. Consider the unit equilateral triangle with the sides representing: (i) the proportion of goods common to both the definition under consideration and the standard, (ii) the analogous proportion of bads and (iii) the GB distance. Each side may be regarded as an axis between 0 and 1 upon which each of the three distance measures may be plotted. Joining the points from each of these measures creates a second triangle within the equilateral triangle. The closer the vertices of the new triangle to the original vertices, the closer the new definition in terms of whichever distance measure that vertex represents.

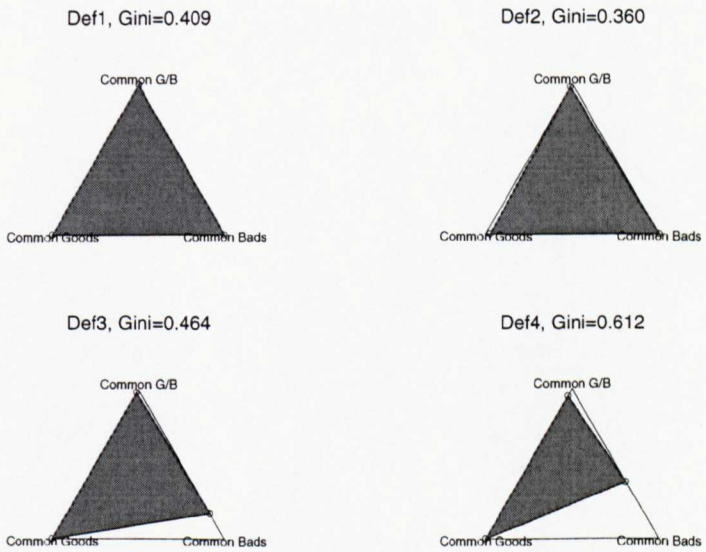

Figure 6.9: *Visual representation of individual definitions*

Figure 6.9 displays the four triangle plots corresponding to definitions 1 to 4. Definition 1 is the standard definition and so the inner and outer triangles are identical. The triangle for definition 2 almost matches the outer triangle yet the Gini coefficient is substantially lower than for Definition 1. Definitions 3 and 4 both produce much improved Gini coefficient. These plots provide an instant visual feel for the closeness of the new definitions. Figure 6.9 clearly conveys that for this data set, classification performance can be enhanced if the definition is altered such that the bad class contains fewer observations.

	Definition 1	**Definition 2**	**Definition 3**	**Definition 4**
Common Goods	1.000	0.999	0.976	1.000
Common Bads	1.000	0.828	1.000	0.622
GB Distance	1.000	0.978	0.979	0.952
Gini coefficient	0.409	0.464	0.360	0.612
% bads	12.290	10.210	14.560	7.470

Table 6.2: *The breakdown of the good/bad comparison between the Definition and the alternatives.*

Notice that the proportion of common bads seems to decrease as Gini coefficient increases.

6.6 Canonical Correlation Analysis

We now turn to a different approach that allows incorporation of definition flexibility. Canonical correlation analysis is a statistical technique which falls into the category of variable reduction. It is a qualitatively different type of model from the logistic approach. Rather than model some predetermined definition as we have done previously, this method allows the data to *decide* the definition.

Credit scoring data often has large numbers of variables. Chapter 2 discusses the two main data sets used in this thesis. They have approximately 100 and 600 variables in total on their respective databases, although only a small proportion are ever used in the modelling phase. The variables may be categorised into two types. Firstly, predictor variables which are used collectively to explain the

relationship between the predictors and outcome so that future performance may be assessed. Secondly, the group of variables which may be incorporated into the bad definition which is to be modelled. Attempting to include too much information can result in models that are difficult to interpret or overfitted. Canonical correlation analysis requires two sets of variables and its aim is to establish linear combinations in each of the variable sets so that the correlation between the linear combinations between variable sets is maximised. See Manly (1994) or Giffins (1985) for details.

Suppose we have two sets of variables $X_1, X_2, \ldots X_k$ and $Y_1, Y_2, \ldots Y_m$. In our case the two groups of variables are those from which the definition is constructed and those used to predict that definition. Canonical correlation analysis involves finding uncorrelated linear combinations of these groups of variables:

$$U_i = a_{i1}X_1 + a_{i2}X_2 + \ldots + a_{ik}X_k$$

and

$$V_i = b_{i1}Y_1 + b_{i2}Y_2 + \ldots + b_{im}Y_m$$

n linear combinations can be obtained, where n is the minimum of k and m. These linear combinations are such that (U_1, V_1) has the largest correlation and (U_n, V_n) the least. In this section we restrict attention to (U_1, V_1), the first canonical variate.

Section 6.3 proposed that a plausible region should be defined so that any definition falling in that region would be acceptable. On this basis the definition yielding the classifier with best performance should be adopted. Here we are suggesting a further step, that is to find the linear combination of variables to be included in the bad definition which are highly correlated with the predictor variables. This canonical correlation approach uses a different kind of bad definition.

Such a radical change leads to some new problems that must be addressed before attempting to construct a model:

1 A decision must be reached on the variables to be incorporated in the bad definition.

2 Variable selection must be performed to determine which predictors are to be used. This can be through expert opinion or an empirical approach.

3 The model must have meaningful interpretation, preferably comparable with that the bank already implements. Basing a model on an optimal linear combination of predictors could lead to a model with nonsensical interpretation. For example, one may obtain a definition that is optimal in terms of predictive power yet not commercially viable.

4 The classes must be based on the value of the first canonical variate. As with error rate a threshold must be chosen which partitions the observations into two classes. In many cases it would be difficult to avoid an arbitrary decision in this situation. However, if company policy is to define some proportion of the applications with the worst predictions for future performance, no problem is caused.

As with any scorecard, care must be taken to ensure that any combination of variables used in model construction is a sensible combination. Models where one variable dominates or models such that all observations are assigned to the same class will be of limited interest.

6.6.1 An Example

The problem is described in Section 6.2. The variables that constitute Definition 1 were used to perform a canonical correlation analysis.

A high value for the correlation of the first canonical variate indicates that more of the relationship between the definition variables and the predictor variables

can be explained by (U_1,V_1). The first canonical variate used in the following analyses has 0.374 correlation.

Partitioning the first canonical variate of each analysis such that the bad class consists of approximately 10% gives the ROC curves in Figure 6.10.

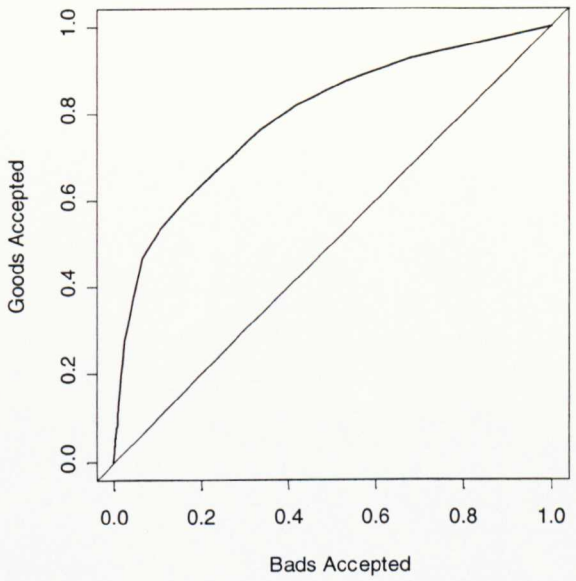

Figure 6.10: *ROC curve from a canonical correlation classifier*

The Gini coefficient for this example is 0.587 – far superior to the model derived from a logistic regression of Definition 1, which give Gini coefficient 0.409.

6.6.2 Model Interpretation and Viability

The canonical correlation model was constructed such that a lower score indicates potentially undesirable accounts.

Table 6.3 shows that large excesses and monthly end balances generate high negative scores, whereas large credit turnover and maximum balance contribute positively to score.

Canonical Coefficients	
Excess	-0.1302
CTO	0.9999
EndBal	-0.4175
MaxBal	0.2714

Table 6.3: *Canonical coefficients*

Definition 1, Section 6.2 states an account is classed as bad if at least one of the following is true:

i) Excess > £500

ii) Excess > £100 and Maximum balance < 0

iii) Credit Turnover < 10% of Monthly end balance

Contrasting the canonical correlation approach with Definition 1:

- In i) large excess would produce a large negative score.
- In ii) moderate excess combined with a negative maximum balance would both result in negative scores.
- In iii) if the monthly end balance is large, it is more likely that an accounts' credit turnover will be less than 10% of the months end balance, large monthly end balance results in a negative score component.

The interpretation of the canonical model generally agrees with the trends implicit in the logistic regression model built with Definition 1.

The canonical correlation model is in some sense an optimal model using these variables, and can readily be interpreted in an analogous way to that of the logistic model, produces superior Gini coefficient (34% larger than the logistic model) and provides an interpretation which is plausible from a business viewpoint.

Chapter 7

Population Drift

7.1 Introduction

In the credit industry scorecards are constructed using large databases of information derived from past accounts. The performance of these classifiers is assessed using an independent test set, which is usually a random sample drawn from the same historical data but which has not been involved in model building. Implicit in this method is that any future sample classified using the scorecard is drawn from the same distribution as the data used to construct the scorecard. However, when the resulting scorecard is used, the population of new applicants may exhibit characteristics that have changed from those upon which the classification rule was constructed. These differences between historic and future populations may result in sub-optimal performance of the classifier. Influences such as economic climate may lead to particular variables systematically changing. Others may be affected by marketing campaigns or drift with time due to product popularity. For example mobile phone ownership is increasing. This may prove to be a predictive variable at present, but in several years time the vast majority of the population may own a mobile phone. If this is the case, predictive power will be lost. Such phenomena do occur but are not generally accounted for in credit models.

Once the classifiers are finalised the resulting scorecards may be used in the marketplace for considerable periods. During the operational lifetime of a

scorecard it is common that no modifications are implemented to account for changes in the population; although it is common that the classification threshold of a scorecard may be altered to offset performance deterioration. If these effects are sufficiently large the performance of a scorecard is likely to deteriorate. Weiss and Indurkhya (1998) acknowledge this problem. When dealing with evolving populations they suggest that the test set should be drawn from the most recent time period. This practice is not commonly acted upon in the credit industry. We propose a model that incorporates effects due to population drift and suggest further ideas for its development.

This chapter comprises three stages. Firstly, unsecured personal loan data will be used to demonstrate that this drifting phenomenon does occur in real applications. Secondly, we explore the effects that this drifting has on classifier performance. Finally, a model that adapts to account for population drift is described.

7.2 Does Population Drift Occur?

The data used in this chapter consists of 92258 unsecured personal loan accounts whose applications were accepted during the period 1 January 1993 to 30 November 1997. This chapter will follow the banking procedure closely so that a meaningful comparison can be made.

Typically a bank may split all the predictor variables into a number of indicator variables (see Chapter 2 for details). These indicator variables are then analysed, often using logistic discrimination which aims to predict the response variable, usually some notion of risk. Following these procedures with loan data set UPL(2), described in Chapter 2, resulted in a scorecard containing 17 indicator variables. Figures 7.1 through 7.5 show how the class proportions for some of these indicator variables vary through time. Applications arising in the period from January 1993 to November 1997 were grouped by week and the arithmetic mean for each variable in the scorecard was calculated. Figures 7.1 to 7.5 show these weekly means plotted against time, with a loess curve (Cleveland, 1979 and

Cleveland and Devlin, 1986) added to aid interpretability. These four graphs illustrate four different patterns present in these data. The trends are: decreasing, suspected policy change, constant and increasing.

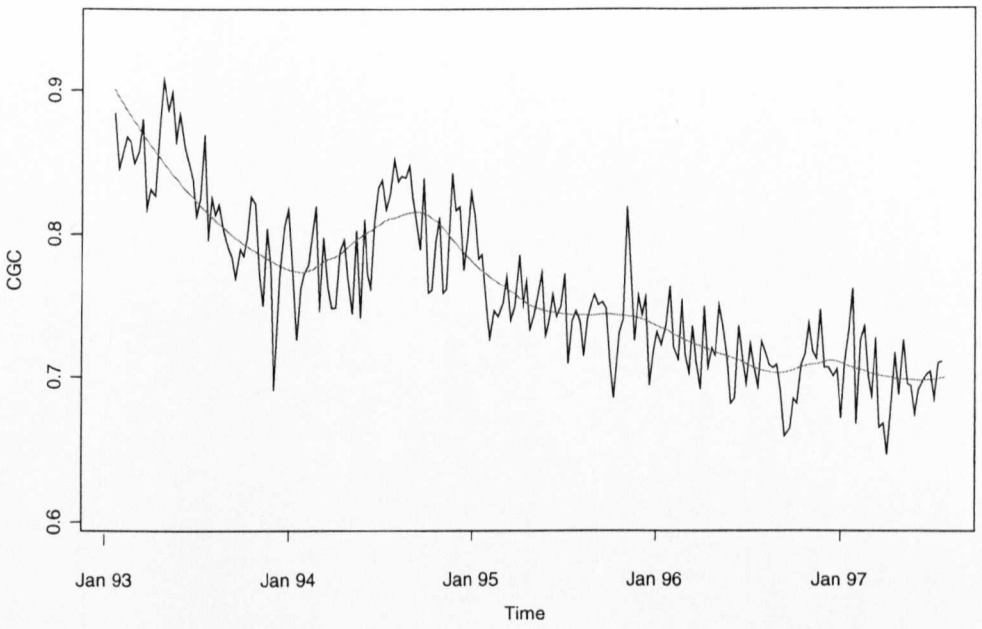

Figure 7.1: *Behaviour with time: Cheque guarantee card.*

Figure 7.1 shows the changing proportion of new applicants accepted each week who have a cheque guarantee card. This variable takes the value 1 if the applicant holds a cheque guarantee card and zero otherwise. Clearly this variable is steadily decreasing with time. This indicates that the proportion of loan applicants accepted with cheque guarantee cards has declined substantially between 1993 and 1997, with a local peak in mid 1994.

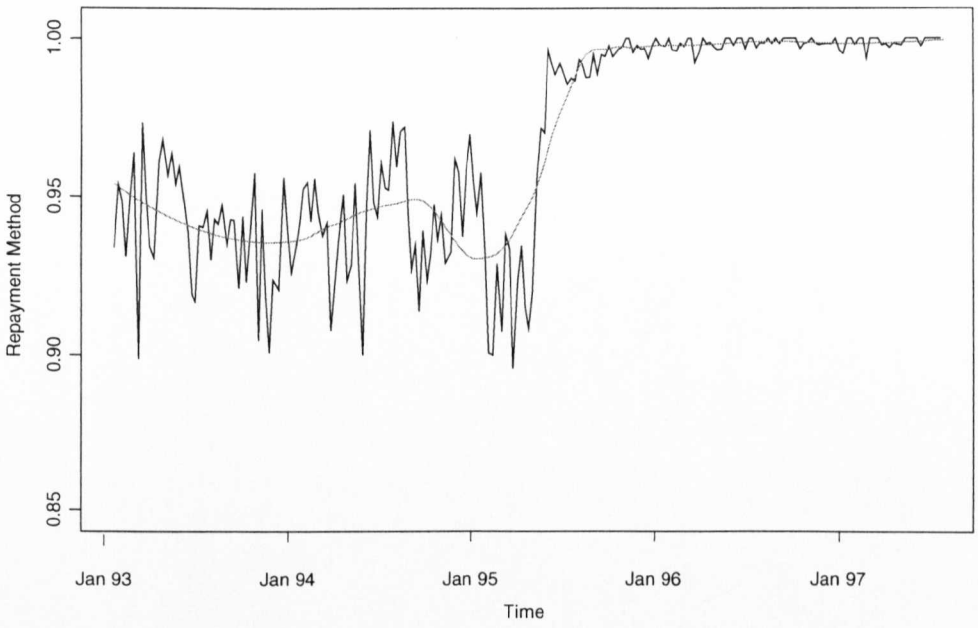

Figure 7.2: *Behaviour with time: Repayment method.*

The variable repayment method takes value 1 if the payment is settled by direct debit, and 0 otherwise. The proportion of accounts repaid by direct debit remained fairly constant for the first three years of the observation period. However, at the beginning of 1995 there is a jump in the trend. A change in bank policy may be one explanation for such a sudden change in the population. The bank may have decided to favour applications whose chosen repayment method was direct debit or, indeed, to enforce this as a condition of the loan. Until 1995, 94% of customers were using a direct debit for monthly repayments, while the rest of the population repaid by monthly payments. From 1995 onwards more than 99% of customers were using the direct debit system. This is a good example of a new innovation, the direct debit system, which has grown tremendously in popularity, in a very short space of time. The result is the predictive power of the variable is reduced.

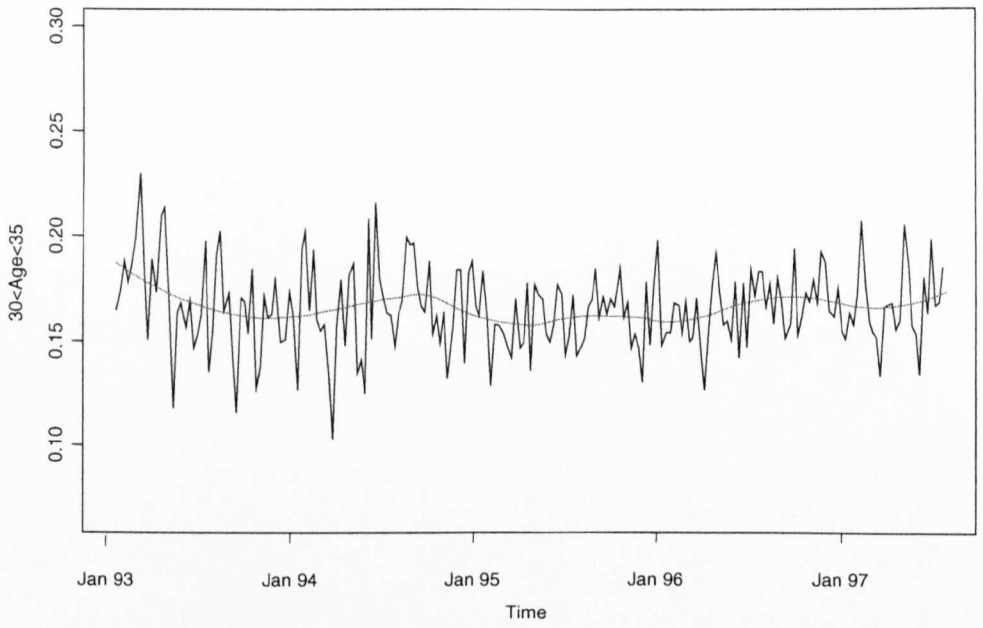

Figure 7.3: *Behaviour with time: An indicator variable which flags customers aged between 30 and 35 years.*

Although age is a continuous variable, common practice involves categorising the variable into several levels. Figure 7.3 represents one level of the categorised version of age. This is entered into the scorecard as an indicator variable taking the value 1 if the age of an applicant is between 30 and 35 years and 0 otherwise. This variable is one of the few that does not show signs of drifting with time. Figure 7.3 does not provide any insight into the evolving age distribution of the applicant population, only between the shift across the two age thresholds (30 and 35) which have been artificially created in the categorising process. The fact that there is not a notable change in indicators of this type simply demonstrates that a similar number of applicants are crossing both thresholds and hence keeping the proportions of zeros and ones in the variable roughly the same. Figure 7.4 shows that there is a definite decrease in the mean age of applicants for the first year of data, after which time the mean age stabilises. This highlights that converting continuous variables to a series of indicators can obscure trends present in the data which may be utilised when constructing a scorecard.

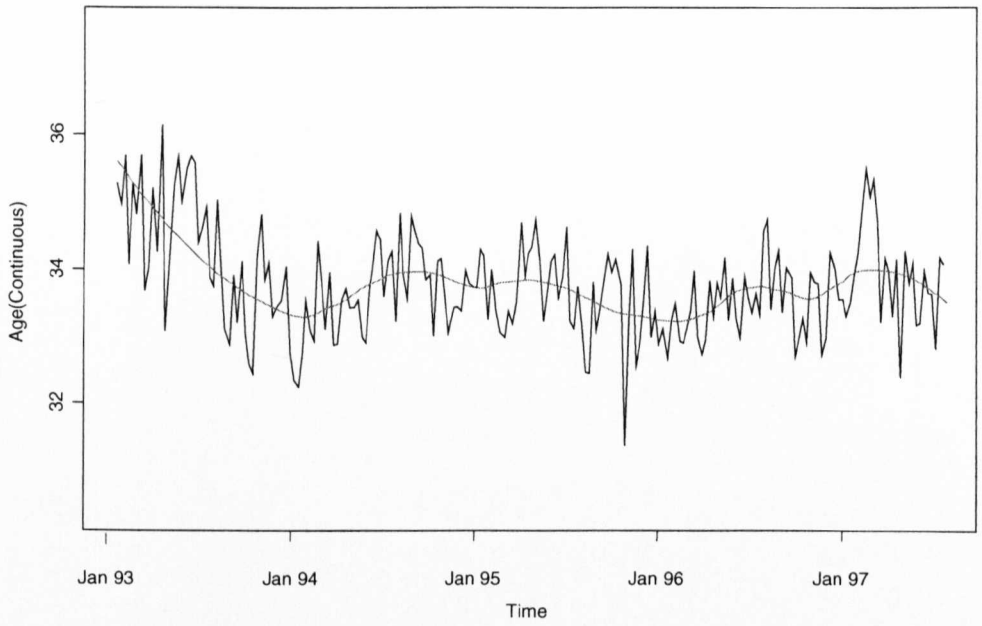

Figure 7.4: *Behaviour with time: Variable - Age, as a continuous variable.*

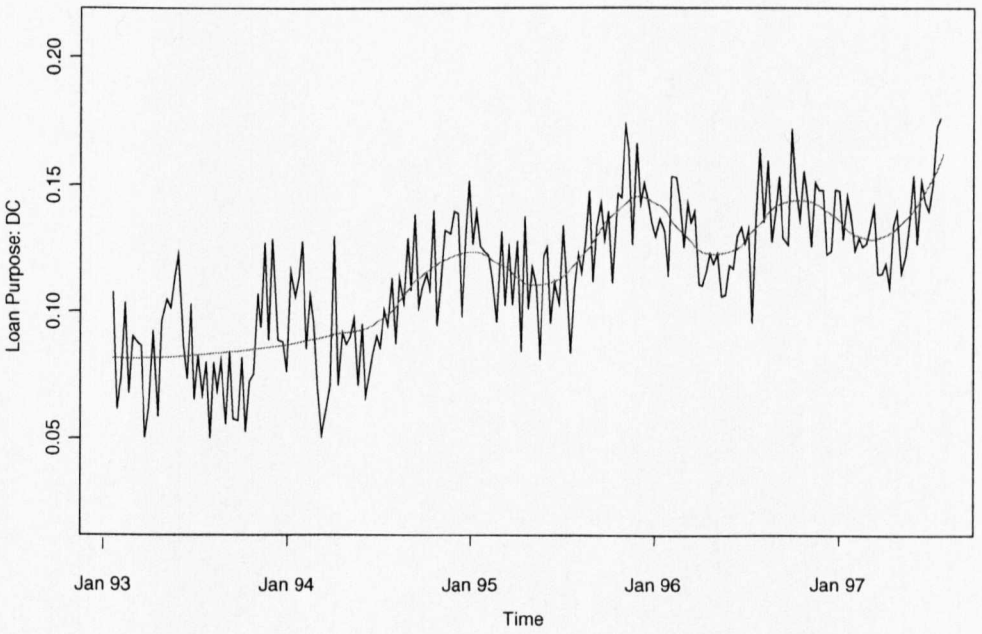

Figure 7.5: *Behaviour with time: Loan purpose - Debt consolidation.*

Figure 7.5 displays the behaviour of the variable that indicates when a loan was used to consolidate other debts. The trend here is increasing with a cyclic effect for the final 4 years. This suggests that a regular pattern is present. Perhaps loans used for debt consolidation reach a peak at certain times during the calendar year. Some trends such as this are not entirely surprising. The economic climate in the early to mid nineties led to many people taking on a large amount of credit. Consequently, a greater proportion of the population encountered repayment problems. Despite at least some of these drifting occurrences being easily explained, current practice is to ignore such population changes until the classification mechanism is replaced.

Note that all these time plots seem to have distinct patterns embedded in their behaviour. The first two years of data seem to exhibit trends quite different from the remaining period. Patterns such as these are virtually impossible to explain in the absence of detailed knowledge of the business and economic history corresponding to this time period. Likely reasons may be changes in the policies used when granting loans, the implementation of a new scorecard or a large change in business operations such as a merger. The bank may be able to derive information that would enable changes in policy to act as the catalyst for changes in their customer base.

7.3 The Impact of Drifting Populations on Scorecards

The previous section of this chapter demonstrated that population drift does occur. Some of the variables that may be included in a scorecard vary dramatically over time. These variations take a wide variety of forms, increasing, decreasing, cyclic, or in the case of a policy change – a step function, or any other combination of these.

Scorecard performance depends upon the predicted probabilities of class membership, $\hat{p}(1|\mathbf{x})$. We have illustrated that drifting effects are present in the populations of the variables used for scorecard construction. However, if the predicted probabilities, $\hat{p}(1|\mathbf{x})$, are independent of these changes then the performance of the scorecards will remain unchanged.

In practice credit scoring takes no account of such drifts whilst the scorecard is in the market place. This may at least be partially due to the history of credit scoring methodology. The credit granting decision has shifted from the obligatory interview with the bank manager whose final decision may even depend on their current disposition, to the automated credit application systems in place today.

With present technology, the development of new credit scoring methods may have followed a different route. Modern computing power would have made it possible to cope with the huge quantities of data that are often sampled prior to analysis. Intensive computational techniques could be used to search for better solutions rather than an 'adequate solution' being implemented. With data storage not posing a problem many variables could have been stored that would enable direct modelling of complex outcomes such as profitability. Traditional techniques have become so ingrained that proposed new approaches often meet considerable resistance from the credit scoring profession.

Common practice involves several steps:

1. Fine tuning the classification rule at the design stage and using historical data in the testing phase to ensure improvements are maintained (Brodley and Smyth, 1996).

2. Using the scorecard in the marketplace to assess the creditworthiness of new applicants.

3. The classification performance is monitored during the operational lifetime of the scorecard. Where possible, any deterioration is compensated for by shifting the classification threshold.

4. When the scorecard is deemed to have deteriorated substantially, a new scorecard is built.

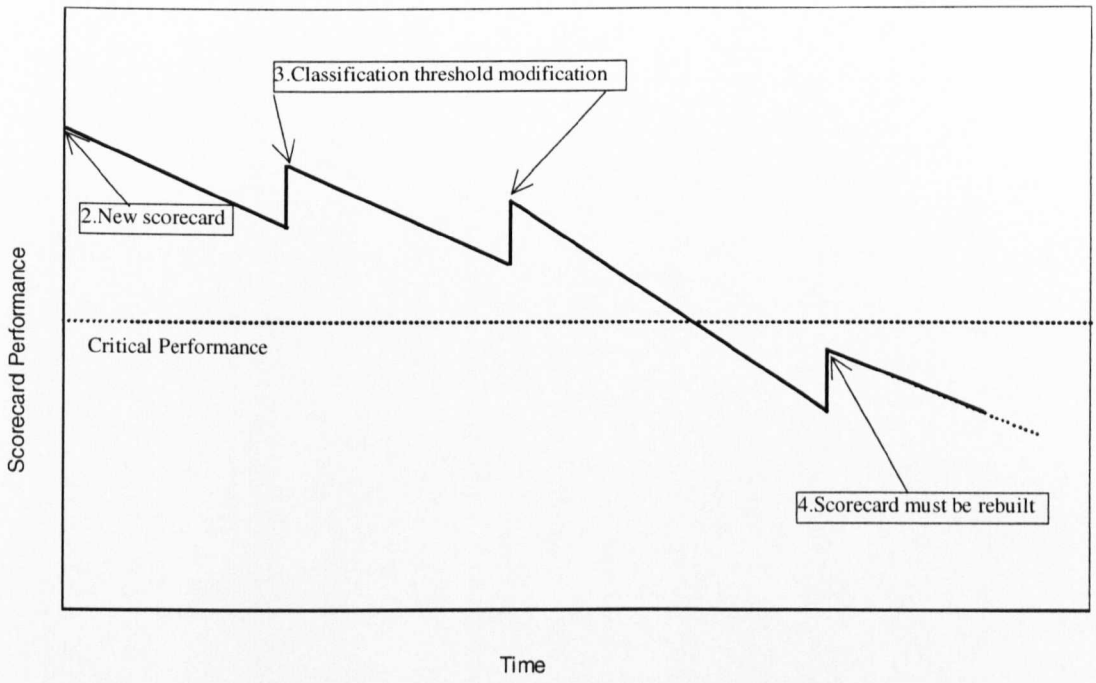

Figure 7.6: *Illustration of a scorecard's deterioration in discriminatory power.*

Figure 7.6 gives a simplistic graphical representation of the above process. In the lifetime of a typical scorecard, the classification threshold may be altered such that the proportion of accepted applications is varied. Crook and Thomas (1992) discuss the degrading performance of the scorecard during the time in which it is used. A classification threshold modification will influence the performance of the scorecard but the precise form of the model remains unchanged. Remodelling entirely, as in point 4, accounts for any population changes by re-estimating the model entirely. A change in classification threshold may help to recoup some of the performance that the scorecard has lost. However, this adjustment does not really address the cause of the problem. The classification threshold is related to the combination of all the variables, hence by varying this threshold, the impact

of each variable in the scorecard is affected. The performance of the scorecard maybe deteriorating due to the drifting nature of just two or three variables. An alternative way of updating the scorecard is presented in Section 7.5.

During the period of observed data we have seen changes in the distributions of these variables. In the next section, we examine effects that these drifting variables have on the classification rules.

7.3.1 Using Data in Yearly Blocks

It is commonplace for a commercial scorecard to be constructed using a whole calendar year's worth of application data split between design and test sets. This procedure seems reasonable in the context of constructing a scorecard that does not take into account the time of year at which an application is made. However, if the population changes over time, it would not be sensible to use either a randomised subset of data from the same period as the design data or to use a sample from a period so wide that performance patterns may be summarised and seasonal trends be overlooked.

In the following section we examine the monthly performance of scorecards which were constructed using UPL data from a whole calendar year.

7.3.2 The Month by Month Performance of a Scorecard

Scorecards are always constructed using historical account information. In this section we construct four scorecards. Each scorecard is constructed using logistic regression, and each estimated using data from years 1993-1996 respectively. We refer to these as scorecard93, scorecard94, and so forth. The performance of these four scorecards was monitored on a monthly basis from the time immediately following its construction until the end of the available data (November 1997).

If drifting populations were resulting in classifier deterioration, a decline in monthly Gini coefficients might be expected. We investigated the Gini

coefficient obtained from each of the four scorecards when classifying observations in monthly groups. Figure 7.7 shows the plot of the monthly Gini coefficients for each of the scorecards. No such trends are obvious. These plots also highlight the variable nature of the performance of the classifiers. The volatility of the performance seems extremely dependent upon the data during the month in which it is applied. A typical month may consist of between 1,500 and 2,500 observations, so it is somewhat surprising that the variation between months is so extreme. The range of Gini coefficients is 0.30 to 0.59 – an enormous difference in an industry where it is common for an improvement in Gini coefficient of 0.02 to be regarded as a classifier enhancement worth implementing. Section 7.4 further discusses this variability.

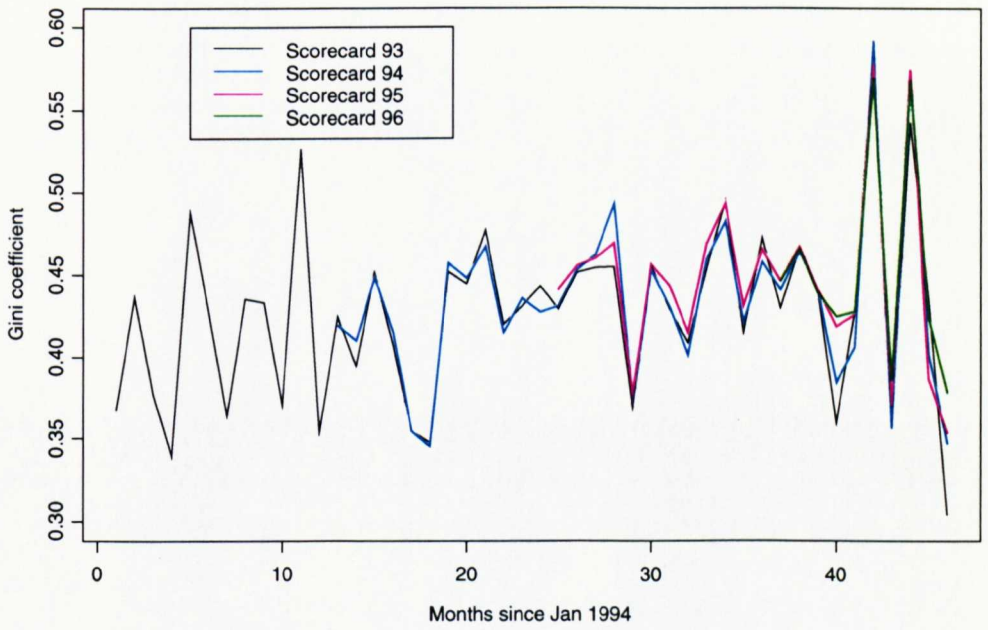

Figure 7.7: *A comparison of the raw performance in Gini coefficient of the four independent classifiers.*

Our primary concern is comparing the classifiers' performance on new credit applications. Figure 7.7 is extremely variable, yet the variability is strikingly similar across classifiers regardless of which of the four independent models is used to classify the data. This variability masks the trends in which we are

interested. Treating the performance of the 1993 classifier as a baseline we can remove the monthly random fluctuations to reveal a clearer picture of any trends changing with time, as in Figure 7.8.

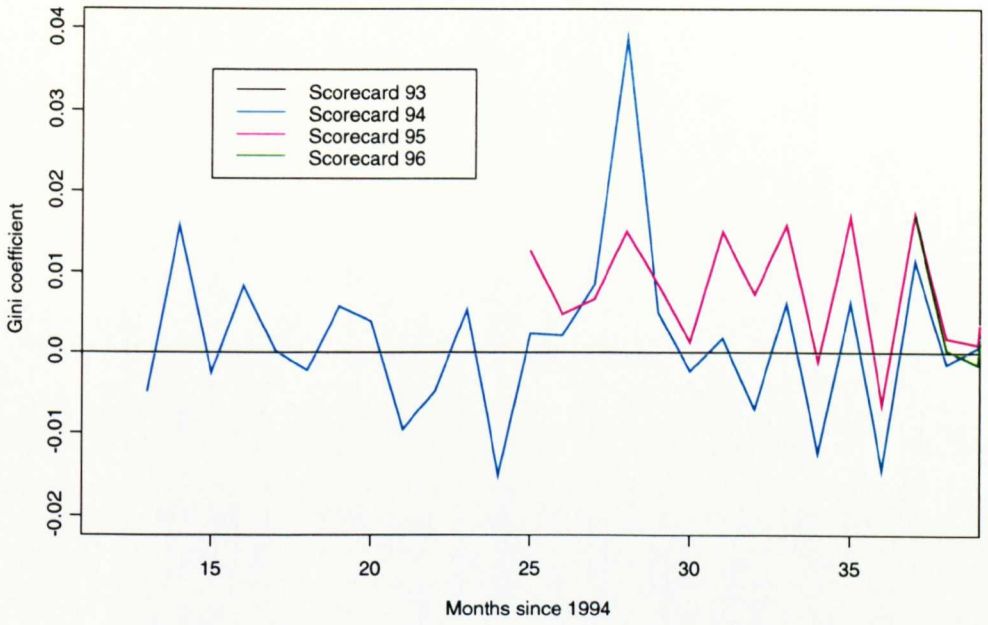

Figure 7.8: *Gini performance of classifiers built using data from 1994, 1995 and 1996 with the performance of classifier 1993 removed as a baseline.*

The cyan line in Figure 7.8 shows the fluctuations in Gini performance for months since 1995 (built using 1994 data) compared to the performance gained by using the classifier constructed from the 1993 data. The purple and red lines are analogous but for the separability of the classifiers built using data from 1995 and 1996.

If drifting populations were causing the classifiers' performance to deteriorate then one would expect this phenomenon to be apparent in the Gini coefficients obtained. However, Figure 7.8 fails to demonstrate that the classifiers built using more recent data are consistently outperforming classifiers that have been in use for considerably longer. It does seem reasonable to suggest that the performance

of the classifier constructed using 1995 data is superior to the 1994 classifier when classifying the 1996 and 1997 data.

Figure 7.9 shows ROC curves for months 36 to 44, corresponding to calendar months January 1997 through October 1997 for the four classifiers. These months were chosen as Gini coefficient variation is more volatile towards the end of the data period, as mentioned previously this is due to the later accounts having less time to default. These results are very similar and reiterate the small differences between classifiers.

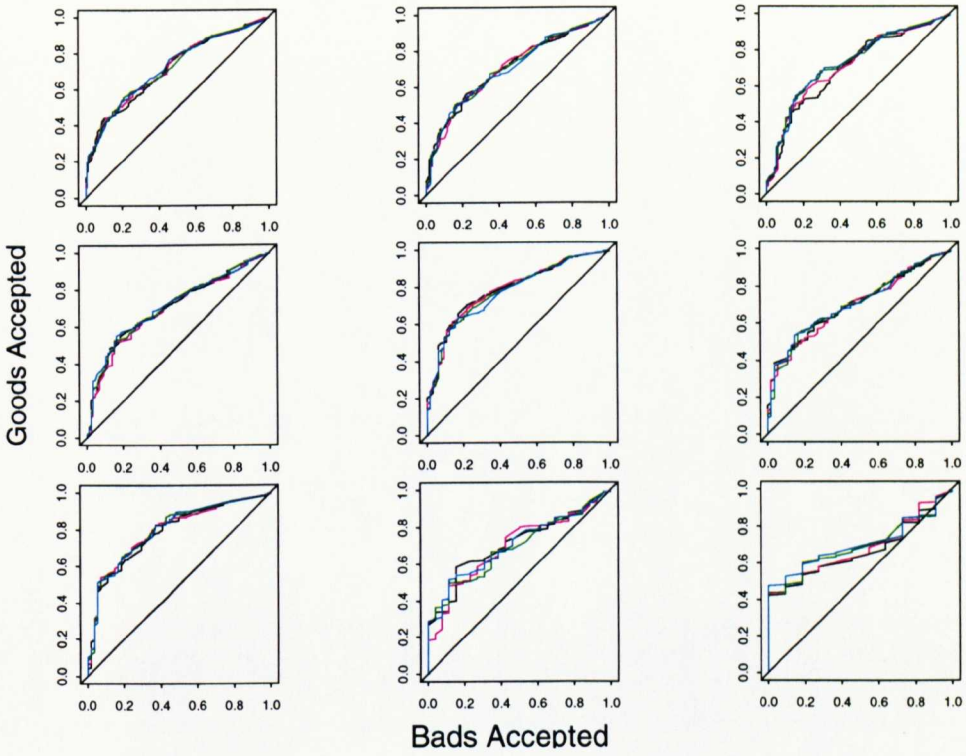

Figure 7.9: *ROC curves corresponding to the monthly performance of the four classifiers. (Legend as Figure 7.8)*

In Chapter 6 we saw that performance, in terms of Gini coefficient, of classifiers was very much driven by the test set. In this situation we see enormous variation in Gini coefficient from month to month, yet four different classifiers, built using data from different time periods provide very similar results. The common factor between these distinct classifiers is the monthly test sets used in assessing their

performance, which suggests that the variation is caused mainly by test set variation.

The implication is that careful monitoring throughout the lifetime of a scorecard is crucial. If checks were only periodical and coincided with a trough on the Gini performance plots, then the impression of deterioration would be incorrect. If a random check coincided with a peak then a scorecard may be left in service when it should have been replaced.

In Section 7.2 we demonstrated that drifting populations are present in the data sets used. The above analysis however, detected only slight deterioration in classifier performance. For this purpose Gini coefficient may not be the most appropriate measure for detecting subtle changes between classifiers. As noted in Section 1.3.2, Gini coefficient summarises across all possible misclassification costs. However in practice when the real costs of misclassification are known only one point on the ROC curve is of interest. Adams and Hand (1998a, 1998b) describe some disadvantages of the Gini coefficient.

An alternative performance measure to the Gini coefficient is misclassification rate. A threshold is required to calculate misclassification rate. This classification threshold may be determined in several ways, two of which are commonly used in the credit industry:

- Policy dictating that all applications having predicted probability below a threshold, decided on the basis of an acceptable level of risk.
- A threshold may be calculated by deciding that the best, 80% say, of all applications would be accepted. That is the 80% which have lowest predicted probabilities of defaulting on their agreements.

Figure 7.10 shows the histograms of the predicted probabilities obtained when using each of the four classification rules constructed using 1993 through 1996 to classify subsequent years of data.

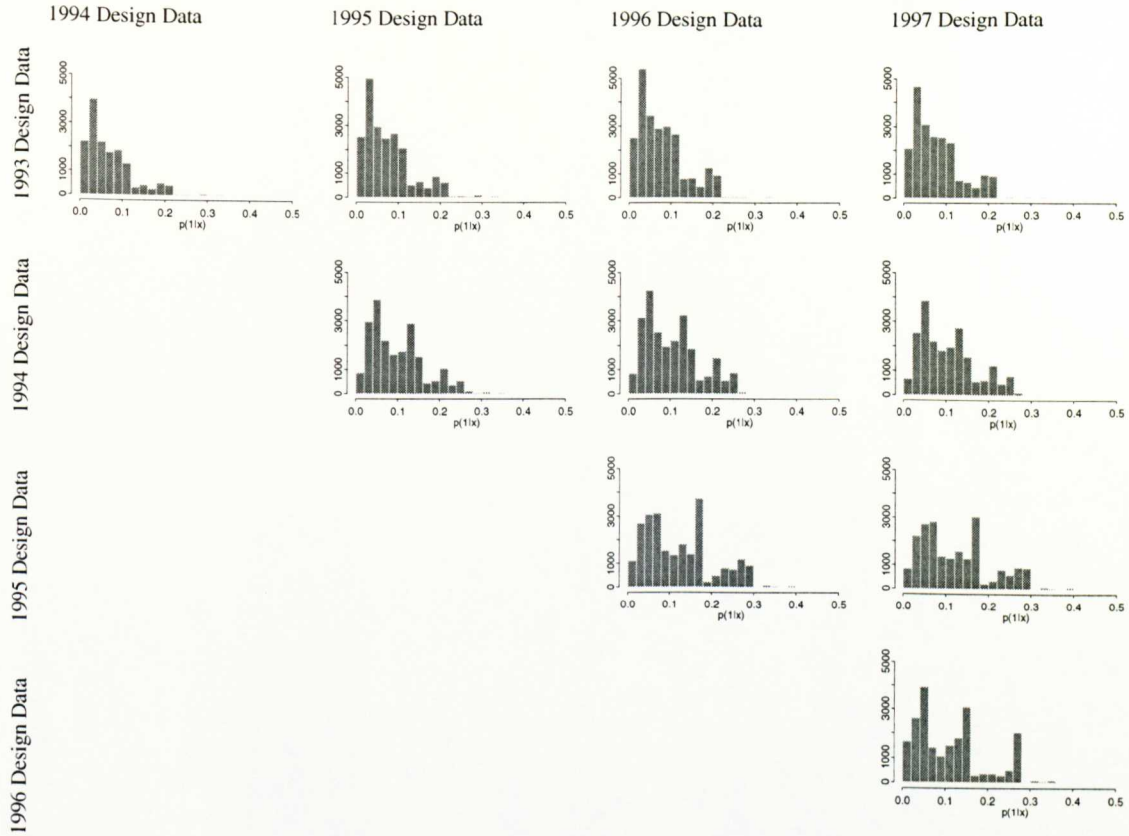

Figure 7.10: *The spread of predicted probabilities by classifier.*

Figure 7.10 displays all histograms of predicted probabilities, $\hat{p}(1|\mathbf{x})$, obtained for each year of data from each classification rule. The columns contain the predictions for the test data 1994 to 1997 with the rows representing the classifiers respectively built using data from 1993 to 1996.

Choosing a threshold for misclassification rate by using the first approach above, that of specifying a threshold probability at the outset gave the plots displayed in Figure 7.11. Probability thresholds of 0.10, 0.15 and 0.2 were used to assign class membership so that misclassification rate may be assessed.

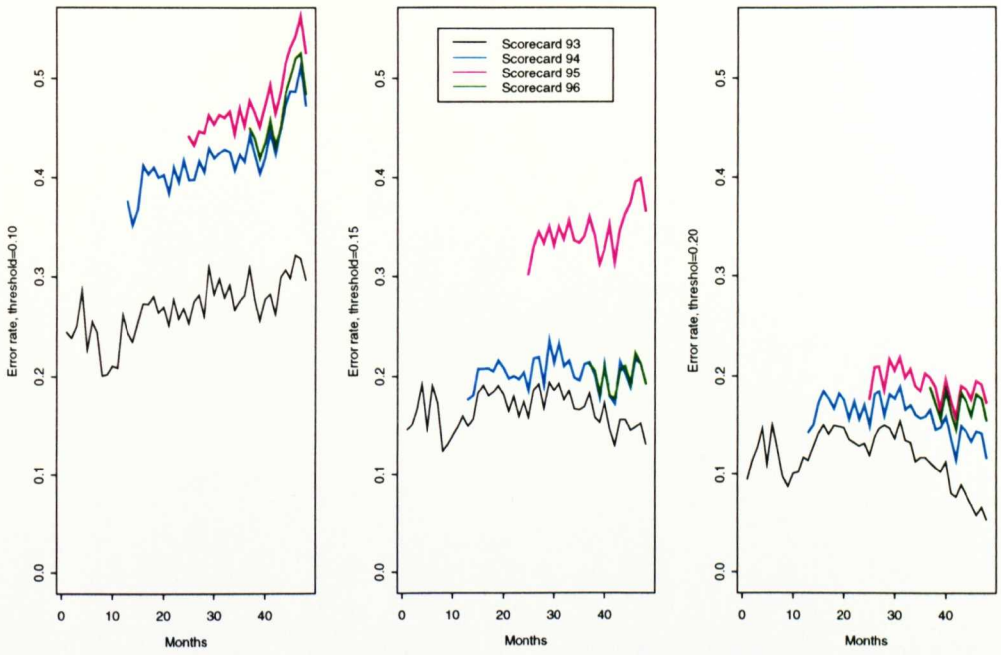

Figure 7.11: *Misclassification performance when the classification threshold is fixed independently of population.*

The plots in Figure 7.11 were obtained by specifying a classification threshold which was to be applied to all four of the classifiers. Fixing the threshold in this way shows that the four classification rules perform very differently, in terms of misclassification rate, depending on the choice of threshold. The priors in the design sets differ considerably. The priors for the bad class in the design sets 93, 94, 95 and 96 respectively are, $p(1)$=0.064, $p(1)$=0.094, $p(1)$=0.117, and $p(1)$=0.110. This may lead to the distribution of predicted probabilities differing substantially between classifiers. If this is the case, a fixed value of classification threshold applied to different classifiers can result in big differences in predicted classes and hence misclassification rates behaving as in Figure 7.11.

A more appropriate choice of threshold may be determined by selecting a value of the predicted probability for which, say 20%, of the predicted population in the test set have greater predicted probability. Figure 7.12 shows this situation.

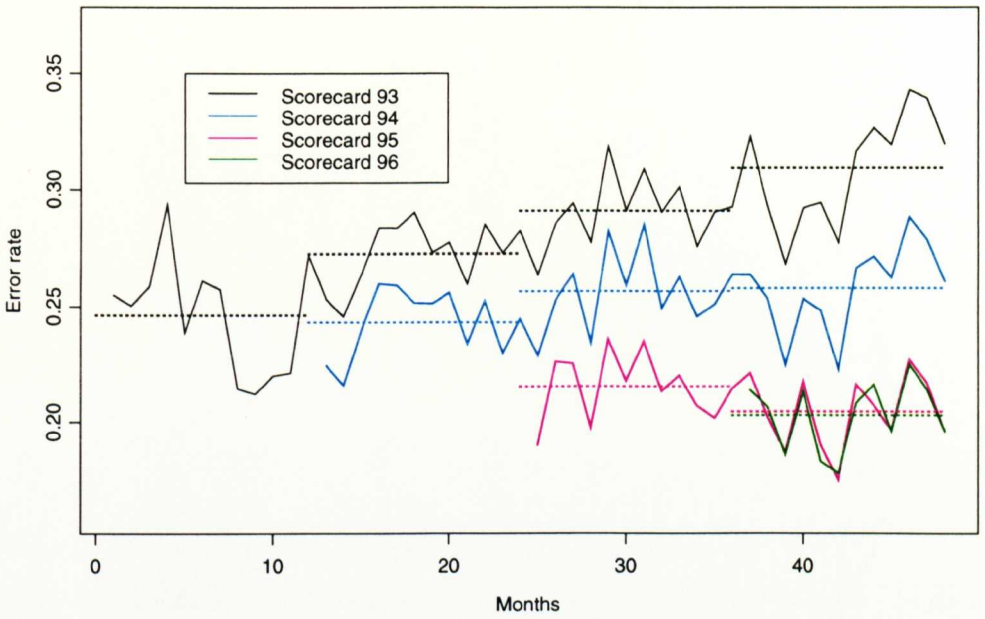

Figure 7.12: *Misclassification rate when the classification threshold is set such that 20% of the population have predicted probability greater than that threshold.*

Figure 7.12 shows the monthly misclassification rate for each of the yearly classifiers 1993 to 1996. The horizontal dotted lines represent the yearly mean misclassification rates. The two classifiers built using the 1993 and 1994 data have misclassification rates that increase with time – that is, their performance is deteriorating. The error rate of the classifier built using the data from 1995 slightly decreases, and for 1996 we have no further data for comparison.

Misclassification rate and Gini coefficient are the most commonly used methods of assessment criteria in the credit industry. The Gini coefficients seem relatively unaffected by population drift, but misclassification rates show large differences.

7.4 Investigation of Monthly Variation of Classifier Performance

Figure 7.7 showed the monthly performance, in terms of Gini coefficient, exhibited by the four classification rules constructed using the data from each of the years 1993 to 1996. The performance is very irregular with no obvious pattern or cause of variation emerging. The analogous results for misclassification rate are presented in Section 7.3, and these too exhibit large variation between months. The aim of this section is to investigate, by means of simulation, possible causes for this variability.

Note that the last eight months are more variable than the preceding months. An explanation may be related to the time at which these accounts were opened. Accounts originating from 1997 are a maximum of 11 months into their loan agreement and so the later in 1997 the account became active the less time it has had to become bad. Consequently, very few bad accounts are present in some months during this time period. Later in this section we demonstrate that the proportion of bads is likely to influence classification performance.

The simulation study is presented in two parts. The first is entirely independent of the real data set, and the second incorporates information extracted from the real data set. The aim of this exercise is to compare the natural variability of standard multivariate normal samples to that of the variability present in the banking data. Specific features of the banking data may be injected into the simulation so that the impact of these can be assessed.

7.4.1 Simulation 1

A simple way in which to simulate a situation comparable with the banking problem is to use two multivariate normal distributions, each with the same number of variables (seventeen) as the banking data, identity covariance matrices and mean vectors zero and m_1. The m_1 is chosen such that the overall separability

of the simulated populations is broadly similar to the separability of the good and bad populations in the real data.

When performing this simulation a classification rule was constructed based on logistic regression using a design sample consisting of ten thousand observations. This classifier was used for all the simulated test sets. The class prior for bads was fixed at 0.1 for the design set and also for all test sets. This provided a reasonable starting point for simulation as the observed monthly priors were between 0.062 and 0.134.

Each simulation run represents a typical month and consisted of 2000 observations. An idea of the variability of the results derived from this simulation is given by Figure 7.13, a histogram of the Gini coefficients obtained from 1000 simulation runs.

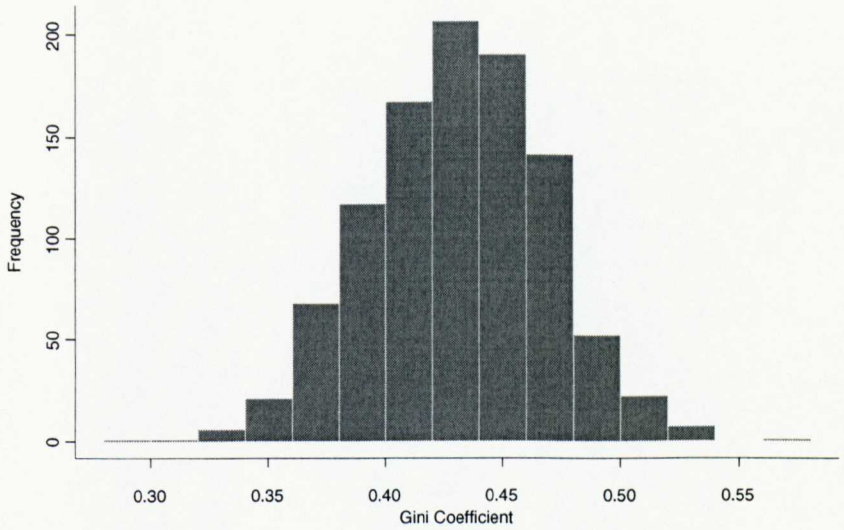

Figure 7.13: *Histogram of Gini coefficients from 1000 simulation runs.*

The observed standard deviation was 0.038 on the 1000 simulations. This is lower than that of the real data, whose standard deviation is 0.045.

7.4.2 Simulation 2

The second simulation differs in one way from the first. In simulation 1 the prior for the smaller class was fixed at 10% of the sample. In this second simulation we use test set priors equal to those in the real monthly data. Only 40 months were simulated compared to the 46 of real data which are available because the last 6 months of accounts in the data set had not been active for sufficiently long that their behaviour could be assumed similar to that of the first 40 months. Observed monthly proportions of bads range from 0.062 to 0.134 in the banking data. Figure 7.14 shows one sample of 40 Gini coefficients resulting from simulations according to this scheme. Visually, the variability of the simulated data is comparable to that of the banking data.

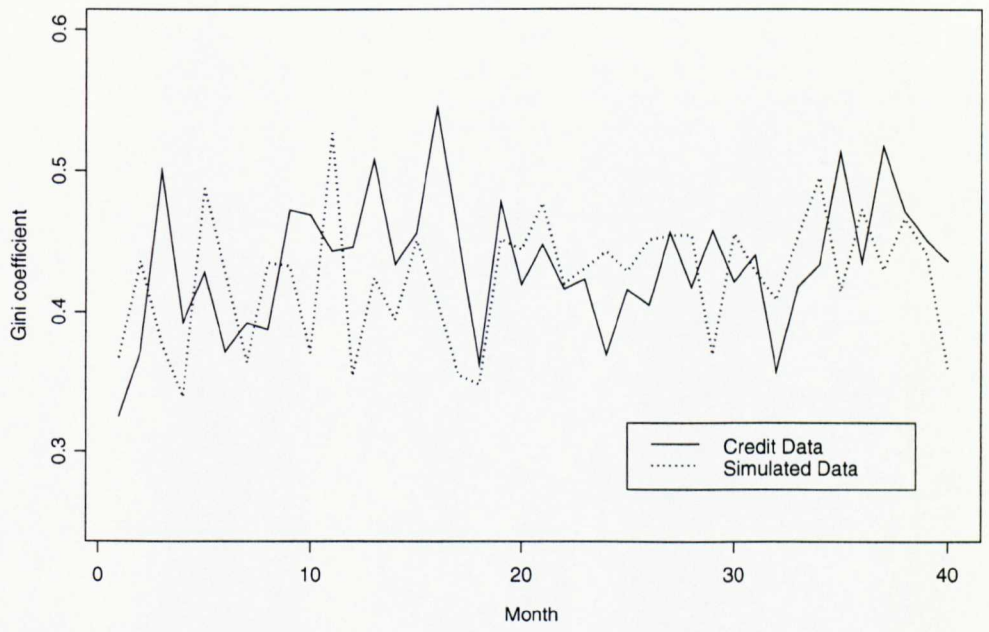

Figure 7.14: *Plot of one simulation run with the monthly credit data results.*

The estimate of variability for this scheme was calculated by computing 500 such 40 sample simulations and calculating the average standard deviation of the 500 obtained. The standard deviation is increased to 0.053 when incorporating

population prior information into the simulation. This is higher than that observed with the real data.

In light of the evidence obtained from these simulations it seems reasonable to attribute the variability of classifier performance to the changing class priors.

7.5 Adaptive models

We have demonstrated that the performance of classifiers built using historical data can be inferior to that of a classification rule constructed using the same variables from a more recent database. Drifting distributions between the times that samples were taken is a probable cause. It is likely that a classification rule constructed using recent data is more able to discriminate between good and bad applications in future samples. In this case it would certainly be beneficial to develop a classification rule that updates as new data becomes available.

The idea of an *adaptive classifier* is not new. Nearest neighbour techniques have been investigated for use in the dynamic updating of models, see Hand (1997) and Taylor, Nakhaeizadeh and Kunisch (1997) for details. Neural networks are also an appropriate method for updating classification rules, see Bishop (1995), Nakhaeizadeh, Taylor and Lanquillon (1998). However, neural network methods involve a considerable computational burden. In the credit literature the idea of continually updating scorecards has been advocated by Bazely (1992). Lucas (1992) provides a discussion of the mechanics of updating. However, little has been published about models that dynamically incorporate new observations.

7.5.1 Incorporating New Observations

Perhaps the simplest way of updating a classifier would be to add extra observations into the model. In this section we will compare the classification results obtained using standard models and those obtained using a dynamic updating approach.

Many practitioners of credit scoring are of the opinion that the more data used, the more accurate the resulting classification rule obtained. Such a statement would be an oversimplification of a complex problem. Others advocate the use of sub-sampling to ensure a reasonable proportion of bad accounts in the sample from which models are constructed. The performance of the model can then be adjusted by correcting for prior information from the whole population (Tarrasenko, 1998). We believe the problem of how to select the data set that will produce the best classification rule is complex and represents a research problem in its own right.

7.5.2 An Adaptive Model

This section proposes a model framework that enables the implementation of an updating classification rule. The updating model is based on linear regression. To facilitate a meaningful comparison, we compare results obtained from the updating model with those from standard linear regression models rather than the logistic regression models we have analysed earlier in this chapter.

Population drift can be described as a linear model whose parameters, β, are continually changing with time, as follows,

$$y_t = \beta_t x + e$$
$$\beta_t = \beta_{t-1} + v,$$

where e and v are independent error terms.

Once the model has been estimated predictions for the behaviour of new applications can be made. Moreover, once the outcome of new applications is known this extra information may be incorporated into the model. The estimates of the regression coefficients at time t have the form $(X^T X)^{-1} X^T Y$. We can update the model at time (t+1) by adding in new points.

When dealing with large data sets re-estimating $(X^T X)^{-1} X^T Y$ each time new observations are to be added to the model can be computationally demanding. No computational problem is caused for the $X^T Y$ factor of the required calculation. However, incorporating a new observation into the $(X^T X)^{-1}$ factor requires a matrix inversion. However, linear regression may be used with the following formula to incorporate an observation into the model without the need for a further matrix inversion.

To add the next point, x_j, into $(X^T X)^{-1}$

$$(X^T X + x_j x_j^T)^{-1} = (X^T X)^{-1} - \frac{(X^T X)^{-1} x_j x_j^T (X^T X)^{-1}}{1 + x_j^T (X^T X)^{-1} x_j} \qquad [1]$$

Using this formula to incorporate new observations, we assess the performance of the scorecard as the classifier evolves.

To illustrate adaptive classification we construct models using data from 1994. The resulting model is used to classify new accounts observed in 1995. We compare this with the classification results obtained from the linear model that updates as the 1995 data is introduced.

Results pertinent to the performance of this model are displayed in Figure 7.15. In terms of Gini coefficient, the adaptive model does not consistently outperform the standard linear regression model, constructed using data from 1994.

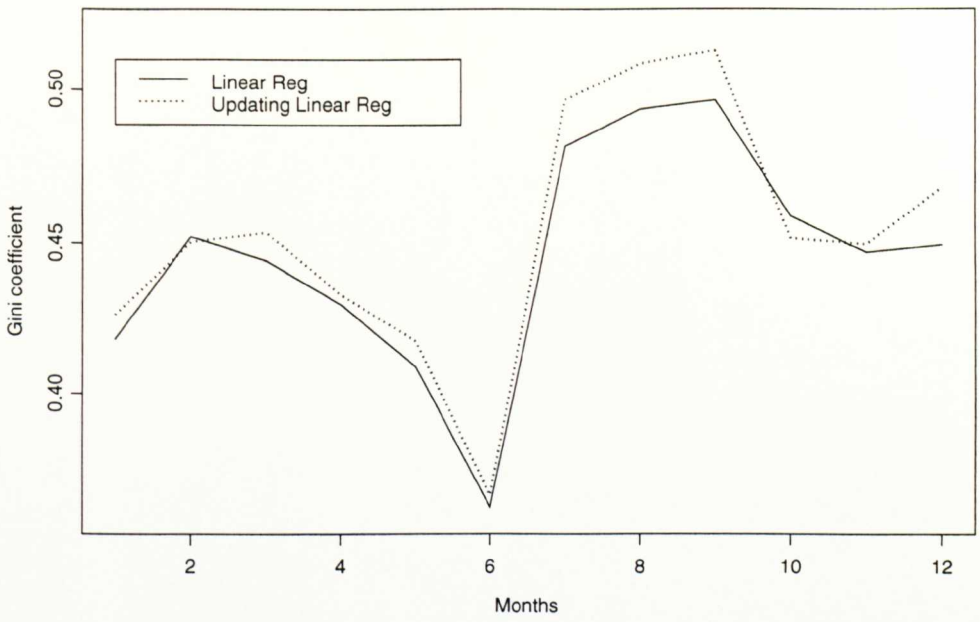

Figure 7.15: *Monthly comparison of Gini performance of standard linear regression techniques and an adaptive classification rule.*

Figure 7.15 shows the results of the adaptive model compared to the classifier derived from linear regression. The starting model for the adaptive classifier was built using data from 1994 and updated using 1995 data.

When adding new observations one would expect change in performance to become more apparent with time. After one month of data has been added to the model, 12/13 of the data is common to the both models. However, once a substantial amount of data was incorporated into the model, we would hope to see improvements in classification performance. We would also expect improvements to increase with time because as the model incorporates more observations the overall model becomes more representative of the current population of applicants. Possible explanations for these indifferent classification results are suggested in Section 7.5.4.

Turning to error rate as the method of assessment, we see results indicating superior performance using linear regression without an updating mechanism, as

illustrated in Figure 7.16. These results are surprising since we have shown that the variables used for prediction are drifting with time and the observations used to update the model are closer (and presumably more representative) to the observations being classified. Figure 7.17 compares the classifiers in terms of bad accounts.

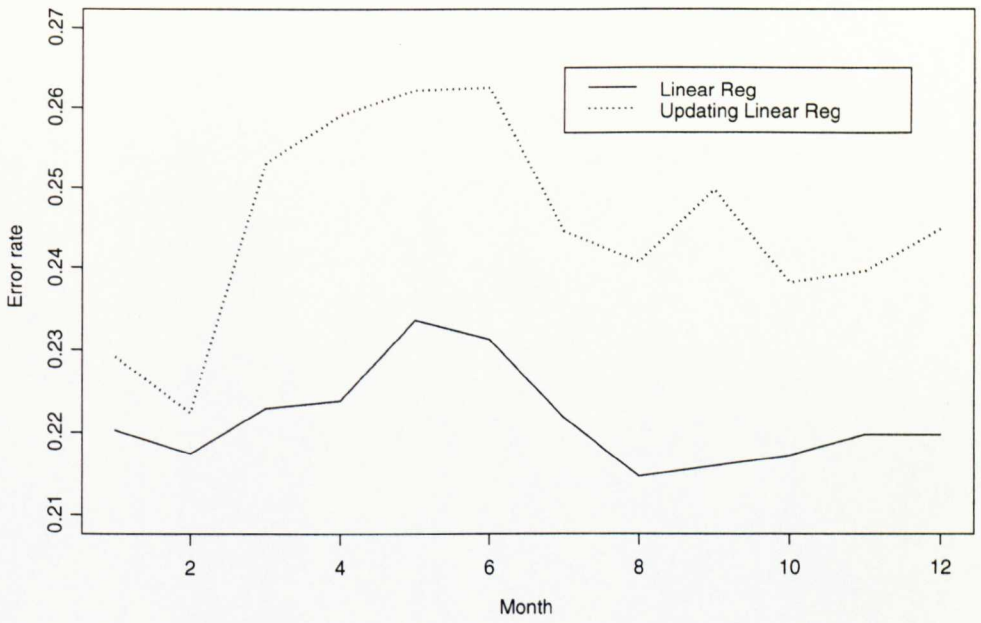

Figure 7.16: *Monthly comparison of the error rate of standard linear regression and an adaptive classification rule.*

Figure 7.17 shows the proportion of bads incorrectly classified for linear regression with and without the updating mechanism. Clearly, a greater proportion of the bad accounts are correctly classified when using the updating method. Misclassification of a bad account is far more costly than the corresponding misclassification of a good account. Consequently, the updating models may be preferred. The apparent rise in error rate can be explained by the adaptive model incorrectly classifying a greater proportion of goods than the standard linear regression classifier.

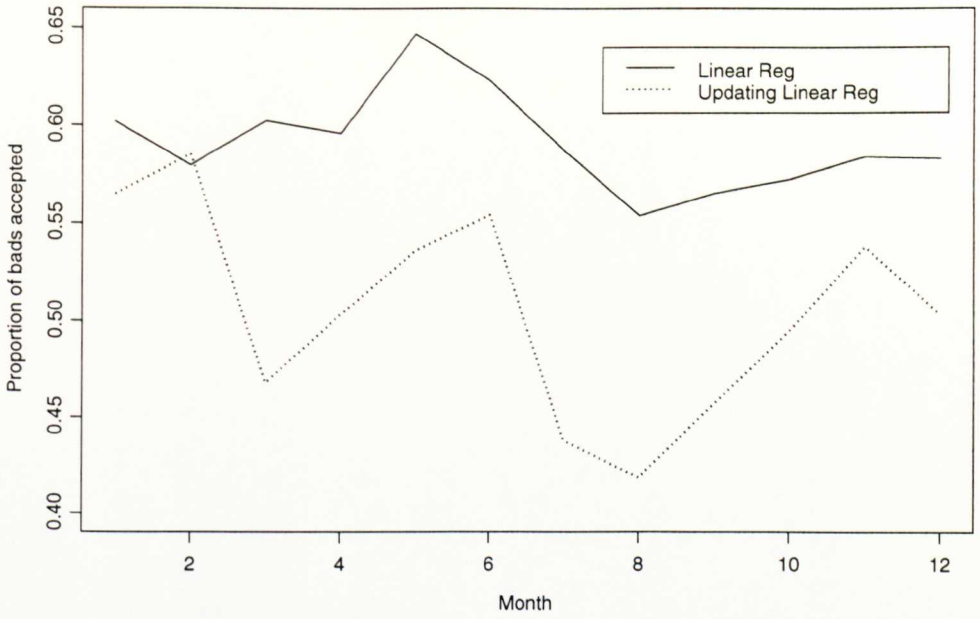

Figure 7.17: *Monthly comparison of bads accepted of standard linear regression and an adaptive classification rule.*

Gini coefficient and error rate are commonly used in the credit industry to compare classification rules. However, the potential loss of a classifier may be its most important feature. In order to calculate the loss (defined in Section 1.4.3) incurred by a classifier we must know the costs associated with each type of misclassification. These misclassification costs are often difficult to obtain. However, useful results may be obtained by working with a likely range of costs. Domain experts stated that the likely cost of incorrectly classifying a bad account as a good is likely to be between 5 and 15 times more costly than wrongly classifying a good as a bad. Figure 7.18 illustrates the loss of the adaptive and standard approaches for three costs in this range.

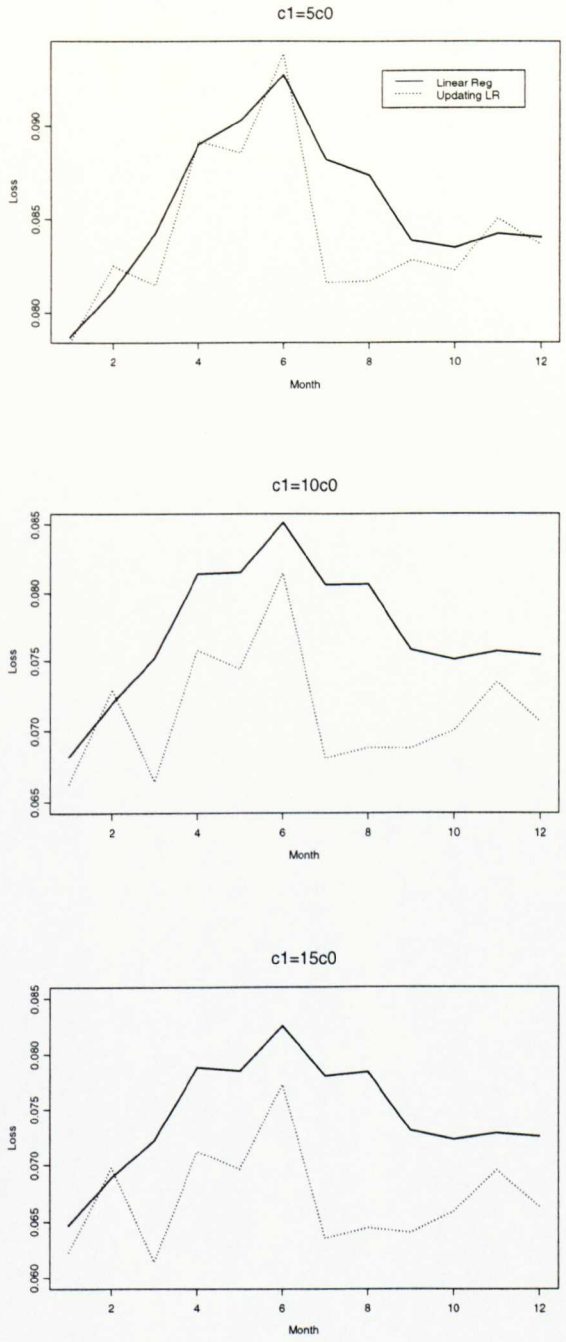

Figure 7.18: *Loss plots of adaptive classifier compared to linear classifier.*

The absolute cost of the updating model is plotted against that of the standard model in Figure 7.18. The different panels in the figure correspond to different ratios of costs. Misclassifying a bad account is deemed 5, 10 and 15 times more serious than misclassifying a good (top, middle and bottom panels respectively).

For cost ratios 10 and 15 the adaptive classifier seems consistently better than the standard linear classifier. As more observations are introduced, the model improves in terms of correctly classifying bad accounts (Figure 7.17). However, the proportion of wrongly classified good accounts in increased.

7.5.3 A Refined Adaptive Model

A possible improvement on the adaptive models investigated in the previous section is to remove the influence of the oldest observations in addition to incorporating new observations into the model. Formula [1] in Section 7.5.2 shows how observations can be incorporated into the regression model without the need for complete recalculation. The influence of a point x_h can similarly be removed using,

$$(X^T X - x_h x_h^T)^{-1} = (X^T X)^{-1} + \frac{(X^T X)^{-1} x_h x_h^T (X^T X)^{-1}}{1 + x_h^T (X^T X)^{-1} x_h}$$

As with the results presented in Section 7.5.2 we illustrate using the model constructed using data from 1994 and update this model using data from 1995. In this section we compare the results obtained by constructing a single model, using linear regression and using that model to classify the following year of data with the updating model (that uses the same linear regression model as starting point). The model is updated as the class of a new observation becomes known. We use error rate and Gini coefficient to compare these methods.

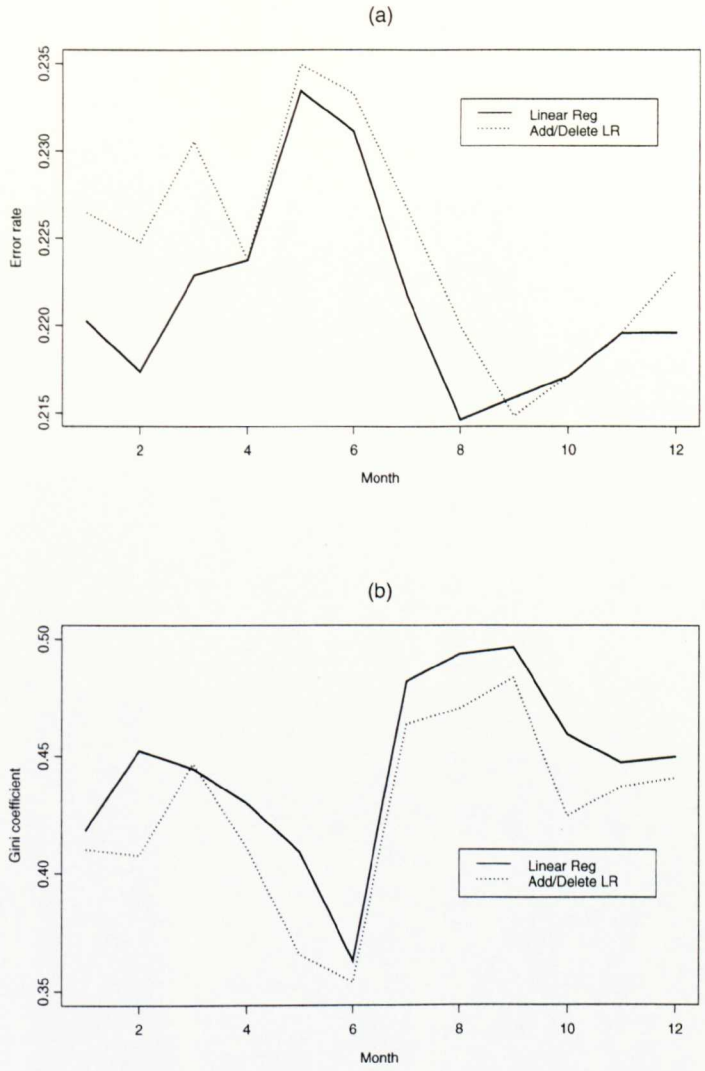

Figure 7.19: *Comparison of linear classifier and adaptive classifier using error rate and Gini coefficient.*

Figure 7.19 shows that the adaptive classifier does not outperform the standard linear regression when an add/delete strategy is used, either in terms of error rate(a) or Gini coefficient(b).

The loss plots analogous to those of Figure 7.18 have been omitted. For each cost ratio investigated it was observed that adaptive classification and standard classification had very similar absolute costs.

7.5.4 Discussion of Adaptive Classification results

Results for the updating models were disappointing. However, in principle this approach seems sensible. We have only briefly assessed one source of data using two updating methods. This section seeks to explain these results and also to suggest ways in which improvements may be made.

We have demonstrated that shifts do occur in the variables used in scorecard construction. Using updating methods ensures that the most recent observations contribute to the construction of the scorecard, i.e. those most representative of the current populations. This alone suggests that a scorecard constructed using an updating rule should have superior accuracy to that of a standard scorecard. Possible reasons for this are given below.

A large amount of observations whose underlying distributions have shifted from the design data, would be required to influence an initial model which was constructed using say, 20,000 observations. The first observation used to update the model contributes 1 in 20,001 to the model, unless earlier observations are removed. However, as updating progresses, the total number of observations increases and the impact of a new observation becomes less. In some sense, the influence of the most recent observations is down weighted.

Surry and Radcliff (1997) suggest benefits are to be gained in using as much data as possible. Their examples utilise up to one million records at the model building stage. However, we have performed extensive model building exercises with the credit data sets used in this thesis. Using various sizes of design set, models were constructed and an independent test set used to assess the performance of each model. Assuming the sample used to construct the model is drawn from the true population of, in our case, unsecured personal loans, the results suggest that there is an upper limit of Gini coefficient. That is, there exists an optimal Gini coefficient for a particular problem regardless of the size of the design sample. This implies that the performance obtained from a design sample of size p, will not be improved upon should a classifier be constructed from a

sample greater than size p, where p is the critical sample size required for optimal performance. Figure 7.20 and 7.21 illustrate some of the results obtained from these experiments.

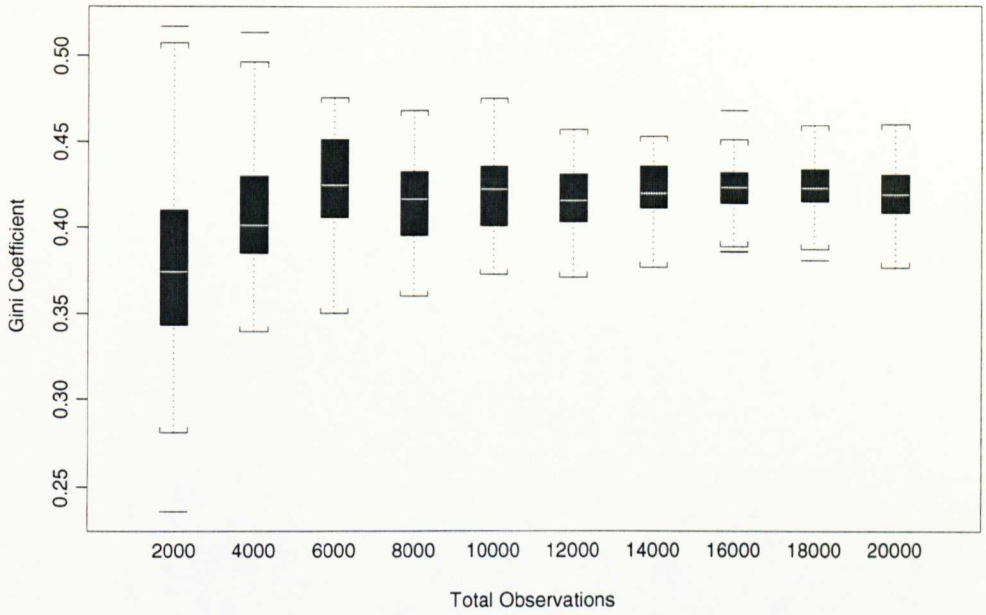

Figure 7.20: *Varying Gini coefficients with increasing sample size.*

One hundred logistic regression models were constructed for each design sample size between 2,000 and 20,000 in increments of 2,000. For each model a random sample was drawn from the design data. Figure 7.20 shows the boxplots of the Gini coefficients obtained from independent test samples. The figure shows that the variation of Gini coefficients for particular sample size is decreasing as the sample size increases. The mean appears to stabilise once the sample size reaches 10,000 observations. Figure 7.21 shows the mean Gini coefficient for each sample size with standard error also plotted. The graph produced by plotting these means seems to be seeking a limit, as expected (Piper, 1992).

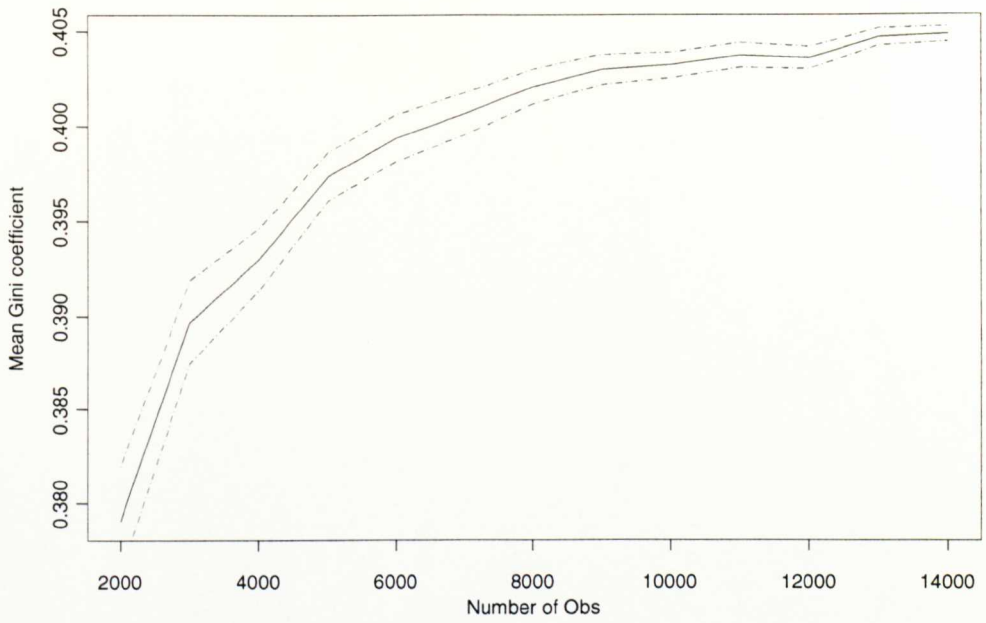

Figure 7.21: *Mean Gini coefficient (with standard error) plotted against design sample size.*

This being the case, we would surely seek to incorporate those observations most representative of the current population of applicants.

Note that the above explanation relies on the samples being drawn from a population having static distributions. However, with only small population drifts apparent this may still provide a reasonable explanation why Gini coefficient remains relatively unchanged. Note, this says nothing about the ability for other assessment criteria to detect small changes in improvement.

The size of the population priors may have a large impact on classifier performance. We observed that error rate is substantially increased when new observations are added without deletion. On the basis of Figure 7.17, an explanation is that many more good accounts are incorrectly assigned. The updating model correctly classified more of the bad accounts. A better approach to help understand the intricacies of these models would be to draw samples of equal size and fixed class priors while preserving the date order. The results

obtained from an updating classifier under these conditions would be free from any effects introduced by the differing samples.

The classifier used in an add/delete updating approach will be dependent on the size of the data set to construct the initial model. This suggests research into the optimal number of observations on which to construct a model may be useful. As a result, updating strategies may be extended such that observations are incorporated (without deletion) until the model is estimated from a data set of a given size. Once this critical size is achieved, the influence of the oldest observation may be removed when new data points are included in the model.

The results presented in this chapter are, of course, based on the two simplest approaches. The first is not ideal as the newest observations have the same impact on the model as the oldest. The second makes use of the most recent m observations and discards the old accounts. However, such a sharp cut-off is a rather artificial situation. All m observations have equal contribution to the model yet those just outside the m provide no contribution whatsoever. A better approach maybe to use a time decaying weighting function.

7.6 Conclusion

Ideas for updating scorecards were proposed in this chapter. Initial results were not encouraging. However, given that drifting effects are apparent in the populations, updating provides a superior framework with which to deal with credit scoring problems of this form. We have suggested several ideas in which further research is required to make progress in this area.

Chapter 8

Conclusions

8.1 Introduction

In this chapter we summarise the work presented in this thesis. The main conclusions are reiterated. We discuss various extensions to the work outlined above and propose various routes for further research. Finally, we make some concluding comments about current practice in credit scoring.

The intention of this thesis is to identify areas in credit scoring where there is potential for large gains rather than simply investigating more and more sophisticated classification techniques, with the aim of 'improving' upon the existing classifier's performance. Many authors such as Henley (1995) and Thomas (1998) have noted that scorecard performance is relatively insensitive to the actual classification rule used. This thesis contains several novel ideas that are directly applicable in commercial credit. These ideas address the credit problem from a different perspective, and tackle situations that may not previously have been identified as areas in which improvements could be made. Adopting this wider viewpoint opens up new areas in which research in credit scoring can be conducted.

8.2 Uncertainty and Change in Credit Scoring

There are many sources of uncertainty in credit scoring. In particular, Chapters 4,5 and 6, highlight uncertainties inherent in many class definitions. The class definitions used in credit scoring are not precise – there is no absolute definition of a bad credit risk. Definitions do vary between different credit products. Institutions offering similar credit products may implement different definitions with which to obtain their predictions. The class definitions used in many credit applications have not been selected in any methodical way. Instead, these definitions have evolved from early intuitive definitions used in the infancy of credit scoring.

Class definitions for consumer credit products essentially fall into two categories. The first category consists of those definitions that refer to a single variable, while the second refers to definitions based on a set of intermediate variables. These types of definition were discussed in Chapters 5 and 6 respectively, and the main points are described below.

8.2.1 Global Models

Global models were introduced in Chapter 5. These models allow a change of class definition without the need to explicitly rebuild the classification rule. Definitions in this category are based on a single variable, used as a proxy for creditworthiness. Examples of such variables used in these definitions are instalments missed in the repayment pattern of an unsecured personal loan, unpaid monthly credit card obligations, and mortgage repayment irregularities. Class definitions for each of these examples are generated by determining the magnitude of the credit 'irregularity' that warrants the label 'bad'. Fixing the size of this irregularity effectively partitions the underlying continuum used as a proxy for creditworthiness. This approach inevitably incorporates some degree of arbitrariness. Credit is not like many scientific applications where there are unquestionable class definitions: credit risk definitions can be varied. Moreover,

external influences such as the economic climate may dictate that the definition should be modified. In these cases a global model can prove to be very valuable because it embodies all possible definitions. When using conventional scorecards, if a change in definition is required then re-estimation is required.

We investigated the impact of changing definitions in Chapter 5 using various real and simulated data sets. We observed that the effects of a change in definition are varied. These changes are problem specific and related to the difference between the original and alternative definitions. Three consequences of a change of definition were observed:

- Classification performance remained relatively unchanged.
- Classification performance would deteriorate unless the existing classifier was modified.
- Classification performance would deteriorate unless a new rule was constructed according to the new definition.

We demonstrated that implementation of a global model would, in some cases, prove invaluable. In addition, we showed that the global model approach would perform at least as well as standard methods in other situations where explicit re-estimation of the classifier could be avoided. We argue that global models implemented in all appropriate situations would provide systems that perform as well as existing solutions, yet are easily adjusted for changes in class definitions. In conclusion, adopting global models in an environment where definitions were subject to change would lead to superior results.

8.2.2 Optimal Models

Chapter 6 describes an example where the class definitions depend upon the values taken by several variables. Parallels may be drawn between this situation and that when a global model is appropriate. This is an extension of the framework in Section 8.2.1, in this case four variables were incorporated in the class definitions. In this case a logical combination of these four partitioned

continua form the definition. We examined the impact of changing each of these partitioning values on classification performance. Plausible limits were obtained for each variable included in the definition. These limits define a region of definitions. Each of these definitions may be regarded as a sensible alternative to the standard.

Given an acceptable region of definitions, huge improvements in performance can be made. In our example Gini coefficients could be improved by 50%. We also proposed measures of 'closeness' (Section 6.5), which can be used to compare alternative definitions with the standard definition. When a comparison to the standard definition is required, it is often the case that extra constraints are imposed. These constraints limit the choice of acceptable definitions. Even under these circumstances we show that improvements in classifier performance can be obtained which are of far greater magnitude than can reasonably be expected from the implementation of more complex classification tools.

8.2.3 Population Drift

In Chapter 7 we established that the population of applicants evolves over time, a phenomenon referred to as population drift. Sections 7.1 and 7.2 demonstrate that many of the predictor variables change over time, and some of these changes are considerable. Typically no action will be taken to combat such changes until a new scorecard is implemented. Section 7.3 shows that classifier performance is extremely variable from month to month, although decreasing trends are evident over time. In this section we observed that error rate is much more sensitive to deteriorating classifier performance than Gini coefficient.

In Section 7.5 we propose adaptive models. These can be used to update the scorecard by incorporating the information from new observations as soon as their class is observed. Utilising a range of plausible cost ratios (provided by a domain expert), we demonstrated (in terms of loss) that the adaptive classifier performs better than standard regression approaches.

We have shown by example that it is possible to update a scorecard dynamically as new observations become available without completely re-estimating the model. This allows the scorecard used in the marketplace to be more representative of the current population of applicants. Consequently, this approach is likely to lead to better predictions than the current policy of leaving a scoring system unchanged until such time as replacement is deemed essential.

8.3 Further Research

Despite the widespread use of credit scoring, research opportunities are still numerous. As shown by Section 1.2, the credit literature is limited. This may partially be due to issues of confidentiality. If research proves beneficial then the body providing the funding for the research is unlikely to share the results through publication. Instead, the 'new technique' will be implemented to try and obtain market advantage. Much of the available literature is concerned with the application of new and more complex classification rules to credit data. This section will highlight other areas for research. In Section 8.3.1, extensions to the work presented in this thesis will be proposed and Section 8.3.2 suggests areas where general research may prove beneficial.

8.3.1 Extensions

Chapter 6 may be regarded as an extension of Chapter 5. Global models provide a framework in which the class definition can be modified and incorporated into the classification rule without the need for remodelling. However, the work on optimal definitions does not allow a definition change (a change in the threshold values taken by any of the definition variables) to be incorporated into the scorecard without the need for the model to be re-estimated. Further work may be directed towards generalising global models to deal with definitions that use several variables so that any of the values taken by any of the variables may be changed without the need for a new classifier to be constructed.

Uncertainty in class definitions described in Chapter 6 was explored using one classification technique, logistic regression. Further work could investigate the possibility of a *classifier effect* in the region of plausible definitions. That is, certain parts of the definition region may be better suited to different types of classification rule.

Canonical correlation analysis was briefly investigated in Section 6.6. Rather than search a predetermined region of definitions this technique takes a different approach. This method determines a definition that is in some sense optimal. Initial investigation provided encouraging results. Further work in this area, and others (such as factor analysis) may provide insight into a new class of definitions that are formulated from the structure embedded in the data.

Chapter 7 described population drift and proposed two methods to combat the problem. The first method incorporated new observations into the model and the second removed the influence of the oldest observation for each new point included. Small improvements in performance were evident despite these being the most simple updating strategies. The form of the updating model presents many possible strategies for further improvements. The next obvious refinement would be to use a time decaying weight function so that the most recent observations included in the model would be the most influential, yet older observations may still contribute. Given the huge imbalance of classes in the population one may investigate the effects of weighting good and bad accounts differently.

In addition to different updating schemes, the work on population drift may be furthered by incorporating time series methods. These may be used to detect useful seasonal patterns, or perhaps to influence marketing campaigns.

8.3.2 General Research Areas

Credit scoring practitioners are continually seeking improvements in the classification performance of their scorecards. Measures such as error rate and

Gini coefficient are used to detect perceived improvements. In the context of misallocation costs, these two measures of performance are opposite extremes. Error rate assumes equal misclassification costs. Gini coefficient is cost independent and summarises classifier performance across all possible classification thresholds. Neither of these assessment criteria are ideal for credit scoring applications. Recent developments have led to substantial improvements in this area, as described in Chapter 1. However, these approaches are still in an early stage of development. Consequently Gini coefficient and error rate are still the dominant measures used in the credit scoring community. We feel further research is necessary in this area so that different classification rules can be compared in a meaningful way.

A typical credit data set usually consists of a good class that is far larger than the bad class. Due to this property of the data, credit practitioners often take a sub sample from the good class and use this together with the entire bad class to construct a classification rule. After the construction phase the predicted probabilities are adjusted according to the sample of the population included in the analysis in order to allow for this sub sampling. It is our opinion that this is not a satisfactory approach and therefore there is scope for improvement in this process.

8.4 Conclusion

In this thesis we have described various types of uncertainty and change associated with credit scoring. Perhaps the ultimate aim of any class definition used in credit scoring is to distinguish between the profitable applications and those which will result in a net loss for the company (this is a generalisation, interest may only be focused on accounts that yield a certain level of profit). In addition, the system designed to carryout this task should be sufficiently robust as to ensure that the desired classification performance is maintained. It is our opinion that looking at credit scoring in different ways, as we have throughout

this thesis, is most likely to result in significant improvements in the predictive accuracy of credit scoring techniques.

In order to gain and sustain competitive advantage in the credit market a radical departure from current methods is required – based on new techniques, recent technological advances, and most importantly, the application of knowledge and understanding obtained in recent years.

Bibliography

Adams, N.M. and Hand, D.J. (1998a) Comparing classifiers when the misallocation costs are uncertain. Technical Report, Department of Statistics, The Open University, UK.

Adams, N.M. and Hand, D.J. (1998b) Improving the practise of classifier performance assessment. Technical Report, Department of Statistics, The Open University, UK.

Albright, H.T. (1997) The use of Genetic algorithms for the stochastic evaluation of consumer credit policy. Presented at *Credit Scoring and Credit Control V, Edinburgh*, September.

Arminger, G., Enache, D. and Bonne, T.l (1997) Analysing credit risk data: a comparison of logistic discrimination, classification tree analysis, and feedforward networks. *Computational Statistics*, **12**, 293–310.

Banasik J., Crook J.N. and Thomas L.C. (1998) Not if but when. Working Paper Series No. 97/4, Credit Research Centre, Department of Business Studies, University of Edinburgh, UK.

Bazley, G. (1992) Profit by the score. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 203–208.

Bishop, C. (1995) *Neural Networks for Pattern Recognition.* Oxford University Press.

Boyle, M., Crook, J.N., Hamilton, R. and Thomas, L.C. (1992) Methods for credit scoring applied to slow payers. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 75–90.

Brodley, C., E. and Smyth, P. (1996) Applying classification algorithms in practice. *Statistics and Computing*, 1997, **7** (1), 45–56

Capon, N. (1982) Credit scoring systems: a critical analysis. *Journal of Marketing*, **46**, 82–91.

Chandler, G.G. and Coffman, J.Y. (1979) A comparative analysis of empirical versus judgmental credit evaluation. *The Journal of Retail Banking*, **1**(2), 15–26.

Chandler, G.C. and Coffman, J.Y. (1983) Applications of performance scoring of accounts receivable management in consumer credit. *Journal of Retail Banking,* **5**(4), 1–10.

Chandler, G.G. and Ewert, D.C. (1976) Discrimination on basis of sex under the equal credit opportunity act. Technical report, Credit Research Center, Purdue University.

Cleveland, W.S. (1979) Robust locally–weighted regression and smoothing scatterplots. *Journal of the American Association*, **74**, 828–836.

Cleveland, W.S. and Devlin, S.J. (1986) Locally weighted regression: An approach to regression analysis by local fitting. *Journal of the American Statistical Association,* **83**(403), 596–610.

Coe, C (1997) Business, prejudice and statistics – a Businessman's perspective. Presented at *Credit Scoring and Credit Control V, Edinburgh*, September.

Collett, D. (1991) Modelling survival data in medical research. Chapman and Hall: London.

Cox, D.R. (1972) Regression models and life tables. *Journal of the Royal Statistical Society Series B,* **34**, 187–220.

Cox, D.R. and Oakes, D. (1984) *Analysis of survival data.* London: Chapman and Hall.

Crook, J.N., Hamilton, R. and Thomas, L.C. (1992) A comparison of discriminators under alternative definitions of credit default. In *Credit Scoring*

and Credit Control, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press, 217–245.

Crook, J.N. and Thomas, L.C. (1992) The degradation of the scorecard over the business cycle. *I M A Journal of Mathematics Applied in Business & Industry*, **4**, 111–123.

Crook, J.N., Thomas, L.C. and Banasik, J. (1995) Does scoring subpopulations make a difference? In *Credit Scoring and Credit Control IV.*

Cyert, R.M., Davidson, H.J. and Thompson, G.L. (1962) Estimation of the allowance for doubtful accounts by Markov chains. *Management Science*, **8**, 287–303.

Cyert, R. M. and Thompson, G.L. (1968) Selecting a portfolio of credit risks by markov chains. *The Journal of Business*, **1**, 34–46.

Desai, V.S., Crook, J.N. and Overstreet, G.A. (1995) A comparison of neural networks and linear scoring models in the credit union environment. In *Credit Scoring and Credit Control IV.*

Devijver, P.A. and Kittler, J. (1982) *Pattern Recognition: a statistical approach.* Englewood Cliffs, N J: Prentice Hall.

Durand, D. (1941) Risk elements in consumer instalment financing. *National Bureau of Economic Research*, New York.

Edelman, D.B. (1992) An application of cluster analysis in credit control. *IMA Journal of Mathematics Applied in Business and Industry*, **4**, 81–87.

Efron, B. (1983) Estimating the error rate of a prediction rule: improvement on cross–validation. *Journal of the American Statistical Association* **78**, 316–331.

Efron, B. and Tibshirani, J. (1993) *An introduction to the bootstrap*, London: Chapman and Hall.

Eisenbeis, R.A. (1978) Problems in applying discriminant analysis in credit scoring models. *Journal of Banking and Finance*, **2**, 205–219.

Feelders A.J, le Loux, A.J.F. and van't Zand, J.W. (1995) Data mining for loan evaluation at ABN AMRO: a case study. In Knowledge Discovery and Data Mining KDD–95.

Fogarty, T.C., Ireson, N.S. and Battles, S.A. (1992) Developing rule based systems for credit card applications from data with the genetic algorithm. *I M A Journal of Mathematics Applied in Business & Industry*, **4**, 53–59.

Freidman, J.H. (1995) Introduction to computational learning and statistical prediction. *Twelfth International Conference on Machine Learning*. Lake Tahoe, California.

Freidman, J.H. (1997) On bias, variance, 0/1 – Loss, and the curse–of–dimensionality *Data Mining and Knowledge Discovery*, 1997, **1** (1), 55–77.

Frydman, H., Altman, E.I and Kao, D–L. (1985) Introducing recursive partitioning for financial classification: the case of financial distress. *The Journal of Finance*, **XL**(1), 269–291.

Frydman, H., Kallberg, J.G. and Kao, D–L. (1985) Testing the adequacy of Markov chains and mover–stayer models as representations of credit behaviour. *Operations Research*, **33**, 1203–1214.

Fukunaga, K. (1990) *Introduction to Statistical Pattern Recognition*, 2nd edition. San Diego: Academic Press.

Fukunaga, K. and Flick, T.E. (1984) An optimal global nearest neighbour metric. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **6**, 314–318.

Gehrlein, W.V. and Wagner, B.J. (1997) A two–stage least cost credit scoring model. *Annals of Operations Research*, **74**, 159–171.

Giffins, R. (1985) *Canonical Analysis: a review with applications in ecology.* Springer–Verlag, Berlin.

Glen J.J. (1997) Integer programming methods for discriminant analysis. Working Paper Series No. 97/3, Credit Research Centre, Department of Business Studies, University of Edinburgh, UK.

Hand, D.J. (1986) Recent advances in error rate estimation. *Pattern Recognition Letters* **4**, 335–346.

Hand, D.J. (1997) *Construction and Assessment of Classification Rules*, Chichester: Wiley

Hand, D.J. (1998) Consumer credit. In D.J.Hand and S.D.Jacka (eds) *Statistics in Finance*. London: Arnold.

Hand, D.J. and Henley, W.E. (1993) Can reject inference ever work? *IMA Journal of Mathematics Applied in Business and Industry*, **5**(1), 45–55.

Hand, D.J. and Henley, W.E. (1994) Inference about rejected cases in discriminant analysis. In *New approaches in classification and data analysis.* ed. E. Diday et al. Springer–Verlag. 292–299.

Hand, D.J. and Henley, W.E. (1997a) Statistical classification methods in consumer credit scoring: a review. *Journal of the Royal Statistical Society*, Series A, **160**, 523–541.

Hand D.J. and Henley W.E. (1997b) Some developments in statistical credit scoring, in *Machine learning and statistics: the interface.* ed. G.Nakhaeizadeh and C.Taylor. Wiley, New York, 221–237.

Hand, D.J. and Kelly, M.G. (1998) Supervised classification when the class definitions are initially unknown. Technical Report, Department of Statistics, The Open University, UK.

Hand, D.J., Li, H.G. and Adams, N.M. (1998) Supervised classification with structured class definitions. Technical Report, Department of Statistics, The Open University, UK.

Hand, D.J., McConway, K.J. and Stanghellini, E. (1997) Graphical models of applicants for credit. *IMA Journal of Mathematics Applied in Business and Industry*, **8**, 143–155.

Hand, D.J., Oliver, J.J. and Lunn, A.D. (1998) Discriminant analysis when the classes arise from a continuum. *Pattern Recognition* **31**(5), 641–650

Henley, W.E. (1995) *Statistical aspects of credit scoring*. Unpublished PhD thesis, The Open University, Milton Keynes, UK.

Henley, W.E. and Hand, D.J. (1997) Construction of a k-nearest neighbour credit scoring system. *IMA Journal of Mathematics Applied in Business and Industry,* **8**, 305-321.

Hopper, M. A. and Lewis, E.M. (1992a) Development and use of credit profit measures for account management. *IMA Journal of Mathematics Applied in Business & Industry,* **4**, 3–17.

Hopper, M. A. and Lewis, E.M. (1992b) Behaviour scoring and adaptive control systems. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press 257–276.

Hsia, D.C. (1978) Credit scoring and the equal credit opportunity act. *The Hastings Law Journal*, **30**, 371–448.

Joanes, D.N. (1993/4) Reject inference applied to logistic regression for credit scoring. *IMA Journal of Mathematics Applied in Business and Industry*, **5**(1), 35–43.

Johnson, R. W. (1992) Legal, social and economic issues in implementing scoring in the US. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 19–32

Kalbfleisch, J.D. and Prentice, R.L. (1980) *The statistical analysis of failure time data*. Wiley: New York.

Kelly, M.G. and Hand, D.J. (1997) Credit scoring with uncertain class definitions. To appear in the *IMA Journal of Mathematics Applied in Business and Industry*.

Kelly, M.G., Hand, D.J. and Adams, N.M. (1998) Defining the goals to optimise data mining performance. *Knowledge Discovery and Data Mining KDD–98*. AAAI. ed. Rakesh Agrawal, Paul Stolorz, Gregory Piatetsky–Shapiro. 234–238.

Khoylou, J. and Stirling, M. (1993) Credit scoring and neural networks. Presented at the *IMA conference on credit scoring and credit control III* at the University of Edinburgh, 8–10 September.

Li, H.G. and Hand, D.J. (1997) Direct versus indirect credit scoring classifications. Presented at *Credit Scoring and Credit Control V, Edinburgh*, September.

Langley, I. (1997) Credit scoring techniques outside credit. Presented at *Credit Scoring and Credit Control V, Edinburgh*, September.

Leonard, K.J. (1988) *Credit scoring via linear logistic models with random parameters*. Ph.D. Dissertation, Department of Decision Sciences and Management Information Systems, Concordia University, Montreal, Canada.

Leonard, K.J. (1993a) A fraud–alert model for credit cards during the authorization process. *IMA Journal of Mathematics Applied in Business and Industry*, **5**(1), 57–62.

Leonard, K.J. (1993b) Detecting credit card fraud using expert systems. *Computers and Industrial Engineering*, **25**, 103–106.

Leonard K.J. (1998) Credit scoring and quality management. In D.J.Hand and S.D. (eds) *Statistics in Finance*. London: Arnold. 105–126.

Lewis, E. M. (1992a) *An Introduction to Credit Scoring*. Athena Press: San Rafael.

Lewis, E. M. (1992b) Credit scoring and credit control from four points of view. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press.

Li, H.G. and Hand, D.J. (1997) Direct versus indirect credit scoring classifications. Presented at *Credit Scoring and Credit Control V, Edinburgh*, September.

Liebman, L. H. (1972) A Markov decision model for selecting optimal credit control policies. Management Science 18 (10), 519–525.

Lovie, A.D. and Lovie, P. (1986) The flat maximum effect and linear scoring models for prediction. *Journal of forecasting*, **5**, 159–186.

Lucas, A. (1992) Updating scorecards: removing the mystique. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 180–197.

Lundy, M. (1992) Cluster analysis in credit scoring. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 91–107.

Manly, B.F.J. (1994) *Multivariate statistical methods: a primer*. Chapman and Hall: London.

Marais, M.L., Patell, J.M. and Wolfson, M.A. (1984) The experimental design of classification models: An application of recursive partitioning and bootstrapping to commercial bank loan classification. *Journal of Accounting Research,* **22**, 87–114.

McCullagh, P. and Nelder, J.A. (1983) *Generalised linear models.* Chapman Hall: London.

Michie, D., Spiegelhalter, D.J. and Taylor, C.C. (1994) *Machine learning, neural and statistical classification.* Ellis Horwood, New York.

Myles, J.P. and Hand, D.J. (1990) The multiclass metric problem in nearest neighbour discrimination rules. *Pattern Recognition.* **23**(11), 1291–1297.

Narain, B. (1992) Survival analysis and the credit granting decision. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 109–122.

Nakhaeizadeh, G., Taylor, C. and Lanquillon, C. (1998) Evaluating usefulness for dynamic classification. *Knowledge Discovery and Data Mining KDD–98.* AAAI. ed. Rakesh Agrawal, Paul Stolorz, Gregory Piatetsky–Shapiro. 87–93.

Oliver, R.M. (1992) The economic value of score–splitting accept–reject policies. *IMA Journal of Mathematics Applied in Business and Industry*, **4**, 35–41.

Orgler, Y.E. (1971) Evaluation of bank consumer loans with credit scoring models. *Journal of Bank Research*, Spring, 31–37.

Overstreet, G.A., Bradley E.L. and Kemp, R.S. (1992) The flat maximum effect and generic linear scoring models: a test. *IMA Journal of Mathematics Applied in Business and Industry*, **4**, 97–110.

Parmar, K.B. and Machin, D. (1995) *Survival analysis*: *a practical approach.* Chichester: Wiley.

Piper, J. (1992) Variability and bias in experimentally measured classifier error rates. *Pattern recognition letters* **13**(10), 685–692.

Provost, F. and Fawcett, T. (1997) Analysis and visualization of classifier performance: comparison under imprecise class and cost distributions. In Knowledge Discovery and Data Mining KDD–97, 43–48.

Reichert, A.K., Cho, C–C. and Wagner, G.M. (1983) An examination of the conceptual issues involved in developing credit–scoring models. *Journal of Business and Economic Statistics*, **1**, 101–114.

Ripley, B.D. (1996) *Pattern recognition and neural networks*. Cambridge University Press.

Salzberg, S.L. (1997) On comparing classifiers: pitfalls to avoid and a recommended approach. *Data Mining and Knowledge Discovery*, **1**, 317–328.

Scallan, G. (1997) Winners and losers: Competition and Experiments. Presented at *Credit Scoring and Credit Control V, Edinburgh*, September.

Scott, M.J.J., Niranjan, M. and Prager R.W. (1998) Parcel: feature subset selection in variable cost domains. Technical report, Engineering Department, University of Cambridge, UK.

Sewart, P. (1997) *Graphical and Longitudinal Models in Credit Analysis*. Unpublished PhD thesis, University of Lancaster.

Sewart, P. and Whittaker, J. (1998) Fitting graphical models to credit scoring data. *IMA Journal of Mathematics Applied in Business and Industry*, **9**, 241-266.

Stanghellini, E., McConway, K.J. and Hand, D.J. (1999) Chain graph for applicants for bank credit. To appear in *Journal of the Royal Statistical Society*, Series C, Applied Statistics.

Surry P.D. and Radcliffe N.J. (1997) Why size does matter in credit scoring. Presented at *Credit Scoring and Credit Control V, Edinburgh*, September.

Tarassenko, L. (1998) *A guide to neural computing applications*. Arnold: London.

Taylor, C.C., Nakhaeizadeh, G. and Kunisch, G. (1997) Statistical aspects of classification in drifting populations. In proceedings of the Sixth International Workshop on Artificial Intelligence and Statistics.

Thomas, L.C. (1992a) Dividing credit card costs fairly. *I M A Journal of Mathematics Applied in Business & Industry*, **4**, 19–33.

Thomas, L.C. (1992b) Financial risk management models. In A. J. and W. F. (Eds.), *Risk: Analysis, Assessment and Management*. Wiley.

Thomas L.C. (1997) Data analysis and data mining for profit scoring: a comparison of possible methods of development. Working Paper Series No. 97/1 Credit Research Centre, Department of Business Studies, University of Edinburgh, UK.

Thomas L.C. (1998) Methodologies for classifying applicants for credit. In D.J.Hand and S.D. (eds) *Statistics in Finance*. London: Arnold. 83–100.

Thomas, L.C., Crook, J.N. and Edelman, D.B. (1992) *Credit Scoring and Credit Control*. Oxford: Clarendon Press.

Titterington, D.M. (1992) Discriminant analysis and related topics. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 53–74.

Toussaint, G.T. (1974) Bibliography on estimation of misclassification. *IEEE Transactions on Information Theory* **20**, 472–479.

Watson, R. (1995) Credit scoring from the borrower's point of view: the consumer credit act and data protection implications. In *Credit Scoring and Credit Control IV.*

Weiss, S.M. and Indurkhya, N. (1998) *Predictive Data Mining: A Practical Guide*. Morgan Kaufmann Publishers, Inc. San Francisco, California.

Wilkie, A.D. (1992) Measures for comparing scoring systems. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 123–138.

Wilkinson, G. (1992) How credit scoring really works. In *Credit Scoring and Credit Control*, ed. L.C.Thomas, J.N.Crook, and D.B.Edelman. Oxford: Clarendon Press. 141–160.

Yobas, M.B., Crook, J.N. and Ross, P. (1997) Credit scoring using neural and evolutionary techniques. Working Paper Series No. 97/2 Credit Research Centre, Department of Business Studies, University of Edinburgh, UK.