If you knew what you were doing, it wouldn't be called research.

-Albert Einstein
A Statistical Framework for Quantifying Adaptive Behavioural Risk for the Banking Industry

Farinaz Farhadieh

Thesis submitted in total fulfillment of the requirements for the degree of professional Doctorate in Philosophy

School of Applied Statistics
Faculty of Life and Social Sciences
Swinburne University of Technology
2011
Abstract

Financial institutions are aware of the importance of a well performing credit scoring model and know that, better scorecards result in substantial monetary savings of millions of dollars. Financial institutions also know that in order to remain competitive in the credit industry, continual improvements of scorecards are necessary. Thus, there has been increasing interest in the application of new classifiers in credit scoring from both practitioners and researchers in the last few decades. The focus in recent years has been on the use of new and innovative techniques to classify customers as risky or non risky with the aim of improving the performance of scorecards.

In this thesis, we concentrate on building a behaviour credit scoring model with a novel approach. We create new variables from data that are easily available and indicate current financial situation of customers and use a mixture of techniques to improve the performance of the model. Our model, named Hybrid SOM based Classification Model is a mixture of classification tree and logistic regression which also considers clustering of customers using Self Organizing Maps. The model uses customers’ transaction history of the past one month and predicts the performance of customers as future good or future bad customers. Transaction history of customers is aggregated to daily time series and the customers are clustered based on these time series and types of transactions. However, clustering is not performed on time series directly. Autoregressive models are fitted to each time series representing customers’ one month transaction and autoregressive parameters are extracted along with moments. Further, percentages of total amount spent on different products are calculated for each customer. Customers are clustered based on the types of products, time series parameters and moments. The top three clusters with the highest percentage of future bad customers are modeled separately. These models are then applied to the whole dataset before the final prediction. The predicted probabilities based on the top three clusters with the highest percentage of future bad customers along with product type expenditure, autoregressive parameters and moments are used for the final prediction. A two stage model is used to predict the probability of risky
customers and the amount of money they owe to the bank. Our model is compared to the conventional logistic model and proven to be superior. Based on the Gini coefficient, the proposed Hybrid SOM based Classification Model, developed within this thesis is categorized as a very good behaviour model.
Acknowledgement

A PhD, by its very nature, is a very individualistic venture. However this venture is impossible to complete without the assistance and support of others. I could not have succeeded without the invaluable support of several.

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr Denny Meyer, for invaluable guidance and support during the period of my candidature. Dr Denny Meyer not only advised and guided me through my journey, but also exposed me to industry to get real life experience in the banking field.

I wish to express my warm and sincere thanks to Graeme Gee who believed in me and provided me with the data and all the necessary resources required. Graeme Gee not only sponsored my PhD but gave me opportunities to work at one of the leading banks of Australia to get experience in the corporate world. I had the pleasure of working with interesting people in different financial institutions. Without such experiences, I would not have been able to carry out this research.

I am grateful to my second supervisor, Dr Paul Bracewell for his ideas and guidance. I would also like to thank Dr Alec Stephenson for his valuable time and support.

I owe my sincere gratitude to my friends who always helped me in difficult situations; special thanks go to Ashutosh Ranjan and Natalia Rajendran for their support and help during my PhD.

I thank my colleague Heath Kitson for his valuable PhD experiences, interesting discussions and programming skills. A special thanks to senior modeling analyst Tawfeeq Hussain for sharing his experience and guidance.

I wish to extend my warmest thanks to Faculty of Life and Social Sciences for their support and friendly attitude. Sincere thanks go to Swinburne University for the financial support and the opportunity provided to me to come to Australia.
I am grateful to Nicruz Hamidi for his patience and loving support.

I would also like to take this opportunity to thank my parents for supporting my educational goals, encouraging and motivating me, and for their unconditional love and prayers.

I cherish my experiences during these never ending years of research.
Dedication

This work is dedicated to my mother. It’s her encouragement which has got me so far. Without her love, support and motivation the completion of this thesis would have not been possible.
Tributes

This research is a project between Offlode Ltd (now part of Dun & Bradstreet) and a leading bank of Australia. Due to confidential restrictions no information about the data could be published. However the methodology is applied to sport field for prediction.

At Dun and Bradstreet the methods applied in this research were applied to scorecards and superiority of the proposed method was highlighted.

Declarations

I declare that this dissertation does not contain

♦ Any material that has been accepted for the award to the candidate of any other degree or diploma

♦ Any material previously published or written by another person except where the reference has been made in the text.

.................................................................

Farinaz Farhadieh

4 July 2011
# Table of Contents

**CHAPTER 1 INTRODUCTION** ................................................................. 1

1.1 Banking culture – Introduction .................................................... 2

1.2 Power of credit card ................................................................. 4

1.3 Credit scoring and credit scores .............................................. 5

1.4 Aim of the research .............................................................. 10

1.5 Overview of the project: ...................................................... 11

1.6 Contribution ...................................................................... 16

1.7 Outline of the thesis ............................................................ 18

**CHAPTER 2 LITERATURE REVIEW** .................................................. 20

2.1 Introduction ........................................................................ 21

2.2 Effect of credit cards on consumer credit .................................. 23

2.3 Overview of credit scoring ................................................ 25

2.4 Behaviour scoring - The theory ...................................... 27

2.5 Credit scoring systems in the past .................................... 31

2.6 Comparison of Methods for Credit Scoring and Hybrid Systems: ....... 45

2.7 Exploratory Approaches for Credit Scoring Data: .................. 49

2.8 Discussion .................................................................... 50

2.9 Limitations of credit scoring: ........................................... 57

**CHAPTER 3 DEVELOPING A CREDIT SCORING SYSTEM** ............... 59

3.1 Introduction .................................................................... 60

3.2 Data processing: .......................................................... 61

3.3 Model Development and Validation .................................... 68

3.4 Performance Measures .................................................. 69

3.5 Model Approval .......................................................... 75

3.6 Summary ...................................................................... 76

**CHAPTER 4 METHODOLOGY** ........................................................... 77

4.1 Introduction .................................................................... 78

4.2 Outline of the studies .................................................... 78

4.3 Clustering: ................................................................. 81
8.5 Suggestions for Future Research ............................................................... 220
8.6 Extensions to Other Contexts ................................................................. 223
8.7 Summary ................................................................................................. 224
REFERENCES ............................................................................................... 228
Appendix A ................................................................................................. 242
Appendix B ................................................................................................. 244
Appendix C ................................................................................................. 249
Appendix D ................................................................................................. 252
Appendix E ................................................................................................. 258
Appendix F ................................................................................................. 260
Appendix G ................................................................................................. 262
List of Figures

Figure 1.1  Number of credit card accounts in Australia from 1995-2010.................. 4
Figure 1.2  Market share for Brands.......................................................................... 5
Figure 1.3  Average balance outstanding per card in Australia.............................. 8
Figure 1.4  Rate of annualized credit growth over the past 20 years..................... 9
Figure 1.5  Average individual monthly changes in total credit outstanding........ 10
Figure 1.6  Steps involve in developing a typical practical scoring method .......... 13
Figure 2.1  Gross balance outstanding & annualized growth on credit cards ......... 24
Figure 3.1  Data partitioning for model building..................................................... 67
Figure 3.2  ROC curves ........................................................................................... 74
Figure 4.1  Steps involved in the two studies......................................................... 81
Figure 4.2  Pruning back to get optimum tree size............................................... 101
Figure 4.3  Feed-forward three layer ANN............................................................ 104
Figure 4.4  Structure of a node with a sigmoid function as limiter....................... 105
Figure 4.5  Different types of activation function.................................................. 106
Figure 4.6  An example of MLP with one hidden layer ....................................... 107
Figure 5.1  Cubic clustering criterion...................................................................... 117
Figure 5.2  Pseudo $t$-squared statistic ................................................................ 118
Figure 5.3  Dendrogram for 15 MC transaction amount series............................ 125
Figure 5.4  Dendrogram based on percentage change in monthly expenditure for five MC codes. ........................................................................................................... 127
Figure 7.1  Procedures applied to the sample dataset ........................................... 157
Figure 7.2  ROC-curve using Stepwise Logistic Regression Model for test data ... 162
Figure 7.3  Clusters Produced by SOM ................................................................ 165
Figure 7.4  ROC Curves for cluster 33 using different techniques....................... 168
Figure 7.5  ROC- Curve for different techniques used on the validation dataset... 172
Figure 7.6  ROC Curve for Hybrid SOM based Classification Model using test data. . ........................................................................................................................................... 176
Figure 7.7  ROC curves for hybrid SOM based & stepwise LR models.......... 177
Figure 7.8  ROC Curve for Hybrid SOM based Classification Model using test data. . ........................................................................................................................................... 198
List of Tables

Table 1.1 Outline of chapter one ................................................................. 2
Table 2.1 Outline of chapter two ............................................................... 21
Table 2.2 Classification results from a study by Gravilowsky and Talley (1981) ....... 39
Table 2.3 Comparison Studies in Credit Scoring ........................................ 46
Table 2.4 Comparison of Credit Scoring Models by Ince et al., (2009) .......... 47
Table 2.5 Probability of Type I and Type II Errors of Different Credit Scoring Models in a Comparison Study by Ince et al. (2009) .................. 48
Table 2.6 Misclassification Rate of Different Credit Scoring Models by Irma & Yap (2010) ................................................................. 48
Table 2.7 Advantages and Disadvantages of Common Credit Scoring Techniques .52
Table 3.1 Outline of chapter three ............................................................. 60
Table 3.2 A 2 x 2 confusion matrix .......................................................... 71
Table 3.3 Evaluation of the quality of classification by the gini coefficient .......... 75
Table 4.1 Outline of chapter four .............................................................. 78
Table 5.1 Outline of chapter five ............................................................... 112
Table 5.2 Means and Standard Deviations for Clusters ......................... 119
Table 5.3 Means for AR (7) Parameters for One Day (Φ1) and Seven Days (Φ7) .. 119
Table 5.4 Monthly Cluster Classifications for Randomly Selected 10 Customers .. 122
Table 5.5 Frequency Table for Each Cluster ............................................. 125
Table 5.6 Percentage Changes in Amount Spent for Each MCG for a Single Month ................................................................................. 127
Table 5.7 Average Monthly Percentage Change in Expenditure by MC Group from June to July and Frequencies for Each Cluster .................. 129
Table 5.8 Monthly Cluster Classification .................................................. 130
Table 6.1 Outline of chapter six ............................................................... 134
Table 6.2 Transactional Data ..................................................................... 136
Table 6.3 Extracted Variable Names and Descriptions Based on Transaction Amount ................................................................................... 138
Table 6.4 Variables Extracted Based on Merchant Category Codes .......... 140
Table 6.5  Total Amount Spent on Each Description During March 2007 ............141
Table 6.6  March Transaction for Customer xxx Provided to us by the bank ..........142
Table 6.7  Percentages of Amounts Spent on Each Description Group ..................143
Table 6.8  Payment Data ...................................................................................143
Table 6.9  Extracted Variables Based on Payment Data ......................................144
Table 6.10  List of Input Variables .......................................................................146
Table 6.11  Delinquency Data ..............................................................................147
Table 6.12  Information Value for Each Variable ................................................150
Table 7.1  Outline of chapter seven .....................................................................153
Table 7.2  Frequency and Percentage of Good/Bad Customers ............................157
Table 7.3  Significant Variables Selected Using Stepwise Logistic Regression ..........158
Table 7.4  Confusion Matrix for Test Data Using Stepwise Logistic Regression .......160
Table 7.5  Top Percentiles for the Estimated Probability of a Future Bad Customer .................................................................161
Table 7.6  Information on the 3 worst clusters produced by SOM with 36 clusters 164
Table 7.7  Characteristics of the Top Three Bad Clusters ......................................166
Table 7.8  Overall Misclassification Rates for Cluster 33 ......................................167
Table 7.9  The Significant Variables Based on Cluster 4, 21 and 33 ......................169
Table 7.10  Misclassification Rate for Final Prediction Using Hybrid SOM Based Classifier ..............................................................................................................171
Table 7.11  Gini Coefficients for Different Techniques ..........................................172
Table 7.12  Significant Variables Using Hybrid SOM based classification Model ...173
Table 7.13  Confusion Matrix for Test Data Using Stepwise Logistic Regression .......174
Table 7.14  Top percentiles for the estimated probability of a future bad customer.175
Table 7.15  Top 10% of customers with highest predicted future bad probabilities.178
Table 7.16  Bottom 10% of customers with lowest predicted future bad probabilities ...
...............................................................................................................................179
Table 7.17  Target Definition and Sample Size for March 2007 Dataset ...............180
Table 7.18  Confusion Matrix for entire dataset using Stepwise Logistic Regression ....
...............................................................................................................................181
Table 7.19  Confusion matrix for entire dataset using hybrid SOM based classification model ..............................................................................................................182
Table 7.20  Target Definition and Sample Size for April 2007 Dataset .................. 184
Table 7.21  Actual & Predicted balances for the top 10 percentiles for expected loss ... ................................................................................................................................. 187
Table 7.22  Target Definition and Sample Size for April 2007 Dataset .................. 188
Table 7.23  Information on the 36 clusters produced by SOM ................................. 190
Table 7.24  Overall Misclassification Rates for Cluster 11 ...................................... 191
Table 7.25  The Significant Variables Based on Cluster 11 and 12 ......................... 192
Table 7.26  Gini Coefficients for Different Techniques ........................................... 195
Table 7.27  Significant Variables Using Hybrid SOM based classification Model... 195
Table 7.28  Confusion Matrix using Hybrid SOM based classification Model .......... 197
Table 7.29  Top Percentiles for the Estimated Probability of a Future Bad Customer... ................................................................................................................................. 197
Chapter 1
INTRODUCTION

The beginning of knowledge is the discovery of something we do not understand – US science fiction novelist Frank Herbert (1920 - 1986).
This chapter introduces the topic and the aim of the research. It also describes the contributions to the literature and to financial institutions. The outline of the chapter is presented in Table 1.1.

Table 1.1  Outline of chapter one

| 1.1. | Banking culture - Introduction |
| 1.2. | Power of credit card |
| 1.3 | Creditworthiness and Credit Scores |
| 1.4 | Aim of the research |
| 1.5. | Overview of the project |
| 1.6 | Contribution |
| 1.7 | Outline of the thesis |

### 1.1 Banking culture – Introduction

Banking culture is extremely complicated on the inside although it looks fairly simple to an outsider. It involves a credit culture which deals with measuring and managing risk and a sales culture which is concerned with making profit, satisfying shareholders and selling the products. Managing such a complex culture is a big challenge. In today’s competitive world, the survival and growth of financial institutions has become critical. Adequately managing risk is a necessary condition for the survival of financial institutions and profitability is essential for growth.

Australia has a strong banking sector. Out of 134 countries, the Global Competitiveness Report 2008-2009 by Porter & Schwab (2008) ranked the soundness of the Australian banking sector 4th in the world. There were 58 authorized banks in Australia by August 2009, among these are the four major banks (Australia and New Zealand Banking Group, Commonwealth Bank of Australia, National Australia Bank and Westpac Banking Corporation), a number of regional banks and 44 foreign-owned banks. According to the latest annual report by the Bank for International Settlements (BIS) the
Australian banking system remains highly profitable by international standards (Long, 2011).

Credit income is one of the main sources of income for banks. Credit is an agreement of future payments for products bought or money borrowed now. Today, the number of credit users has dramatically increased and the banks are offering many products to attract more customers. For banks the ability to offer credit has important sales and profit implications but this activity involves a high risk (credit risk), for both lender and borrower.

Three major components of risk in banks and financial institution are Credit risk, Market risk and Operational risk. Amongst all the risks involved in banking, credit risk is the main concern for many bank authorities since it can result in bank failure if not well managed. Credit risk is the risk due to the counterparty to a transaction defaulting on their obligation. In such cases the bank has to bear the loss. The bank is responsible for having enough capital to overcome these losses.

Credit risk is a multi-faceted topic. Consumer credit is credit obtained by an individual for purchases other than property (Guardia, 2002, p. 2). Consumer credit is a term used for all kinds of installment and non-installment individual credit. Credit cards are an example of installment credit. This thesis broadly focuses on credit risk in regard to customer's credit card behaviour. Since this type of credit is unsecured, the bank does not have any sort of collateral to take possession of in the event of delinquency, increasing the amount of capital that is needed to cover this type of credit risk. For this reason, banks should be very careful in their lending decisions in regard to credit cards, selecting only the “right” borrowers.

---

1 Installment credit is repaid in two or more payments.

2 Non-installment credit is repaid in a lump sum. An example of non-installment credit is single payment loans by financial institutions.
1.2 Power of credit card

Credit makes it easy to obtain anything we need or want at any convenient time. But also credit makes it harder to resist buying anything you wish for. For this reason credit can be a burden instead of a convenience. However credit cards have become a requirement in today’s technology driven world. Hiring cars, shopping online and ordering products are much easier if one has a credit card.

The economic growth of Credit Card Companies in Australia reflects their growing importance in our society. In Australia the number of credit card accounts has risen from 6.5 million in 1995 to 13.5 million in 2009 (www.rba.gov.au). By May 2010, there were 14.6 million credit and charge cards issued in Australia. Figure 1.1 shows the number of credit card accounts for the last 6 years in Australia.

![Figure 1.1 Number of credit card accounts in Australia from 1995-2010](www.rba.com.au)

A charge card provides an alternative payment to cash when making purchases in which the issuer and the cardholder go into an agreement that the debt incurred on the charge account will be paid in full and by due date, usually every thirty days, or will be subjected to strict late fees and limitations on card use (Wikipedia).
There were 1.51 billion credit card transactions in 2009, with a total value of $225.8 billion. The number of transactions grew by 3.9% over the past year, while the value of transactions grew by 2.5%. This is the lowest annual growth rate on record since 1989, which can be attributed to the Global Financial Crisis. Credit Cards have a big future as the world becomes less of a “cash in hand” society. In today’s day to day living Credit Cards have become an essential part of everyday life, whether used for petrol, groceries or travel.

Currently MasterCard and Visa are the two biggest credit companies in Australia, sharing 85.6% of the market by March 2010 according to reserve bank of Australia (www.rba.com.au) as shown in the Figure1.2.

![Market share for Brands](source: www.rba.com.au)

*Figure 1.2* Market share for Brands

### 1.3 Credit scoring and credit scores

Looking at the sales culture in banking, it is necessary for banks to attract new customers and retain valuable customers. To do so banks need to target customers who meet certain profitability criteria. In particular banks are interested in choosing customers who are never going to default. If banks are too strict in choosing the right customer, the result will be loss of profit opportunities and if banks are too relaxed in accepting customers, they will cause large unredeemable credits. Therefore, banks need to be very careful in choosing the right customers. However, selecting the right customers now does not guarantee that these customers will not default in the future.
Banks are constantly checking customer behaviour after credit has been allowed in order to understand the current level of risk for each customer. To measure the risk, the banks apply statistical analysis called credit scoring to help them make credit decisions. The result of credit scoring is numerical measures known as credit scores. A credit score is a numerical value that reflects the likelihood of an individual repaying their debt sometime in the future. There are many different credit scorecard systems with differing ranges for their resulting credit scores. A high credit score represents less risk implying that the individual has an increased likelihood of repaying the debt. A customer with a high credit score is known as a good customer. A low credit score represents high risk and implies that the individual has an increased likelihood of not repaying their debt in the future. A customer with a low credit score is referred to as a bad customer. According to the Fair Isaac Company, over 90% of credit card lenders use credit scores when making their lending decisions. Credit scoring is not limited to banks. Credit scores are used by insurance companies and mobile phone companies, employers, government departments, landlords; and their use continues to expand.

Broadly speaking, banks deal with two different types of customers requiring different types of decisions.

- New customers: Should the new application for credit be granted?
- Existing customers: Should the agency grant the request of an old customer to increase their credit limit? How risky are the existing customers? What products to offer to the existing customers to maximize the profit?

The techniques used to determine the answer to the first question are called application scoring whereas the answer to the second question requires behavioural scoring.

Application scoring is more common in the literature to the point that when credit scoring is discussed, one automatically thinks of application scoring. The literature available on credit scoring is scarce due to the sensitivity of credit data. Literature on behaviour scoring is almost nonexistent. In this thesis credit scoring refers to both application and behaviour scoring, however, we shall be concentrating on behaviour scoring.
When a new customer applies for a credit card, the financial institution checks the creditworthiness (risk of lending) of the customer based on characteristics such as age, income, marital status, previous credit history, etc. to see if the credit card should be granted. This process is mainly done through credit bureaus. A credit bureau is a company that gathers information from various sources and provides credit information on an individual’s borrowing and bill paying habits. A credit bureau, in some cases the financial institution itself, analyses the customers’ creditworthiness using statistical analysis. The statistical analysis results in the creation of scores, called application scores.

The statistical analysis for the creation of credit scores is unique to each credit bureau. In today’s competitive world the bureaus can no longer rely only on financial measures. Credit scores now represent an overview of all personal data which is gathered in databases. Based on the database and some unique statistical formula, the credit bureau creates a 3 digit numeric number which provides the client with a measure of their customer’s creditworthiness.

Credit scores are not only used to determine whether credit should be approved for an applicant, but are also used in the setting of credit limits. If the applicant has a score greater than the required score, a credit card will be issued to the applicant. The limit for this credit card is set based on the risk level of the applicant as reflected in their credit score.

Although there is a credit check before the credit card is issued to a new applicant, this does not guarantee that the applicant will keep up their payments in the future. Financial institutions regularly monitor the behaviour of existing customers and calculate the likelihood of customers defaulting in the future. The customers transactional and payment behaviour is monitored and a behaviour score is calculated using statistical analysis. Based on this score, good (not-risky) and bad (risky) customers can be identified helping banks to make good decisions about increasing credit card limits for existing customers or offering new products to the customers. Financial institutions also use credit scores to come up with appropriate strategies to minimize their losses/
maximize their profit. For example they may offer a good customer a higher credit limit in the hope of future financial benefits to the bank.

In today’s world credit amounts are growing steadily because the amount owed by individuals is increasing and the number of people with credit cards continues to grow (Thomas, Edelman, & Crook, 2002). This makes it more difficult for financial institutions to manage their credit risk, but it also provides large amounts of data which can be used to improve the reliability of credit scores. Figure 1.3 shows how the average balance outstanding per card has changed in the Australian banking sector in the last five years.

![Figure 1.3](source: http://www.bankers.asn.au)

**Figure 1.3** Average balance outstanding per card in Australia.

At the end of May 2010, the average balance outstanding per card was $3,249, an increase of 5.0% per card over the past year. The global financial crisis really started to show its effects in the middle of 2007 and into 2008. In ‘The Sunday Morning Herald’ the Melbourne Institute research fellow Dr Edda Claus said in a statement that in the June quarter 2010, for the first time since November 2006, credit card debt is the most common form of debt among Australian households, rather than mortgage debt.

The financial crisis of 2008 resulted in the collapse of large financial institutions, the rescue of banks by national governments, and decline in stock market prices around the world. Many economists believe that this crisis is the worst financial crisis since the great depression of the 1930s. The financial crisis of 2008 contributed to the failure of
key businesses, decline in consumer wealth estimated in the hundreds of trillions of U.S. dollars, substantial financial commitments incurred by governments, and a significant decline in economic activity. Many experts have suggested several causes. Many solutions have been implemented or are under consideration, but still the world economy is under significant risk. Some critics argued that investors and credit bureaus did not accurately price the risk involved with financial products related to mortgages, and that governments failed to adjust their regulatory practices to deal with financial markets during the 21st century.

This financial crisis had a big impact on credit growth. Figure 1.4 and 1.5 shows the rate of credit growth over the past 20 years, and the individual monthly changes in total credit outstanding respectively.

![Credit growth - Total (%annual)](http://www.bankers.asn.au)

*Figure 1.4* Rate of annualized credit growth over the past 20 years

During 2009-2010, annual growth rates have been at their lowest levels since 1993. From July 2009 to July 2010, total credit growth was 2.8%. The highest recent credit growth was seen in 2007 (16.3%), followed by 24 consecutive months of falling annual growth rates up to December 2009.
Throughout most of 2009, monthly growth rates for individual’s total credit outstanding were around the lowest levels since 1992. In 2009, on average, total monthly credit grew by 0.1%. Monthly growth increased to 0.5% during the first few months of 2010 followed by a fall thereafter.

Deterioration in the financial circumstances of customers is often associated with default. If this change in financial circumstances can be detected early, banks are able to ensure that they have sufficient capital to cover this risk. The main purpose of this thesis is to identify changes in the financial circumstances of existing customers as early as possible. This research strives to identify customers who are likely to miss payments in the future.

1.4 Aim of the research

The field of credit risk is mature. Consequently, with any attempt to construct a methodology, the approach in this research needs to be novel not only to satisfy the academic requirements of new research but also to produce better results for the banking industry. In discussions with one of the leading banks of Australia, an improved understanding of retail credit risk has been identified as a priority.
In this project risk behaviour of credit card customers will be modeled at an individual level to predict longer-term changes in behaviour before any activity damaging to the bank has occurred. The primary goal of this project is to identify, customers who are likely to miss 3 payments (bad / risky customers) in the immediate future. The intention is to enable the bank to conduct pre-emptive action minimizing both reputational and monetary losses.

Current systems are based on a 12 month view of historical customer behaviour and assume that a customer’s future behaviour is similar to past behaviour. The intention of this thesis is to develop an early warning system with a historical one month view that fully utilizes all available transactional information collected in the past month and predicts the future performance of customers based on this current information. This time frame (one month past view) can give us a better understanding of a customer’s present financial status.

The Global Financial Crisis had a great impact on many individuals. A customer could have had a good financial situation for many years but could have quickly ended up in trouble due to the financial crisis. If we score a customer’s credit worthiness based on 12 months of data, we cannot capture recent changes in his/ her financial situations. But, with our past month view, this recent change can be captured. The challenge is to minimize the false positive rate for default detection whilst maintaining an adequate capture rate. This will enable cost-effective contact strategies to be conducted in order to better manage defaulting customers. Further analysis of customer response to these contact strategies could enable the customers most likely to benefit/accept assistance from the bank to be identified and prioritized.

1.5 Overview of the project:

Behaviour scoring has become an important task in the credit industry. Behaviour scoring has many benefits including closer monitoring of existing accounts, reductions in credit analysis costs, faster credit decisions and prioritizing credit collections (Brill, 1998).
Despite widespread literature on credit scoring, most of the studies deal with benchmark datasets and, as mentioned previously, few consider behavioural scoring. Few studies are conducted on real-world datasets containing only the most relevant variables and there is a lack of studies in the literature that explain the entire process of scorecard development from the raw data to the final model. There is a need to fill this gap in order to differentiate between the theoretical and practical aspects of scorecard development and to ensure the successful implementation of new approaches.

Behaviour scoring models aim to group customers that share similar behaviour patterns. Using these patterns, banks target different groups to promote new products, increase credit limits, target the groups which will be encouraged to spend more and also come up with strategies to manage recovery if a customer’s repayment ability turns bad. In behaviour scoring models, historical transaction behaviour and payments are considered assuming that the customer’s behaviour will be similar in the future. In order to model a customer’s behaviour, behaviour scoring models establish an association between input variables and an output score, which measures the probability of default. Based on these associations, a score is assigned to each customer and customers are clustered into groups for marketing purposes.

The typical practical scoring method usually involves the steps shown in Figure 1.6. Banks refer to such methods as scorecards, but this is a throwback to the past when manual procedures involving cards were used for credit scoring.
All these steps are essential in developing a scoring method; however, in this thesis the focus is on the ‘data preparation’ and ‘model development and validation’ steps. These three steps have particular potential for improving the performance of behaviour scoring models. For example, different feature selection techniques can be used to choose input variables, different classification algorithms can be used to train models. Various model algorithms can be used with different input variables to see which gives the best result. The choice of modeling objective is the primary key to developing scorecards since it defines a full set of technical estimation procedures that are used to select the “best” model under the objective and also it defines how we assess its validity. In other words the aim of the project has to be very clear before developing a scorecard so that the right techniques are selected according to what is needed.

‘Data preparation’ and ‘variable selection’ steps are very important in credit scoring. It has been found in the literature that the use of new and more predictive variables can improve the performance of scoring methods (Hand & Henley, 1997). Most scoring studies use benchmark datasets which are already clean and contain few variables. Unfortunately, only a few studies have examined the use of variable selection
techniques in behaviour scoring. This is because it is difficult to obtain large real world behaviour scoring datasets containing many variables, and the studies which are based on real world behaviour scoring data utilize clean datasets which do not require variable selection.

The ‘model development and validation’ step is used to discriminate between ‘good’ and ‘bad’ applicants. The better the classifier, the better will be the performance of the scoring method. In this research we apply a selection of different techniques. We compare their performances and use a mixture of these techniques. The purpose of combining techniques in scoring is to produce a final model that is better than the individual models in terms of prediction accuracy (Lee & Jung, 1999/2000). Only a few studies have used hybrid models/mixture of techniques for this purpose (Bahrammirzaee, Rajabzadeh Ghatari, Ahmadi, & Madani, 2011; Chuang & Huang, 2011; Gopalakrishnan, Sridhar, & Krishnamurthy, 1995; Lee & Chen, 2005; Lee, Chiu, Lu, & Chen, 2002).

The data for this research comes from two important Australasian banks. There were several discussions with bank analysts to ensure that this research was relevant and easy to implement within existing “scorecard” development processes.

In this thesis we look at the credit card transaction behaviour of customers to gain a better understanding of their transactional behaviour, we model this behaviour and predict the probability of default. According to the Basel 2 definition “a default takes place when the obligor is past due more than 90 days on any credit obligation”. A defaulter is a customer who defaults on credit. In financial vocabulary a defaulter is referred to as a “bad” or a “delinquent” customer. In this research we follow the Basel definition of delinquent customer and refer to a delinquent customer as a “bad” customer throughout this thesis. A bad customer is a customer who doesn’t pay the minimum expected amount of repayment for 3 months (90 days). A “good” customer is a customer who repays regularly and does not carry the minimum expected payment to the following month. In this research we refer to a customer who has only missed one payment as a “slow-paying” customer. A risky customer is a customer who may become a bad customer in future and a non-risky customer is a future good customer.
Two datasets containing different types of information were analysed in this research to gain more understanding of customers’ transactional behaviour. Two different research studies (i.e., a pilot and a main study) were performed using these two datasets. Both the research studies aim at providing an understanding of the current financial circumstances of customers and predicting future performance based on current financial situation of customers.

**Pilot Research study:** The dataset used for this study was limited to only transaction amounts and type of transaction (for example, Coles supermarkets, Harvey Norman). This pilot research study was useful in getting a better understanding of transaction behaviour. In this study customers are grouped (clustered) based on transaction amounts and the type of product spend for a month. A more sophisticated method of clustering based on extracted global characteristics from time series is also applied to the data to extract useful information. This process is repeated separately for 6 months and customers are clustered in each of these months. The movements of customers between the clusters are monitored to gain an understanding of the changes in customers’ behaviour over time.

**Main Research study:** Large datasets for the above variables and delinquency status are used for the main study. Input variables are created based on the methods used in pilot study. Future performance of customers is predicted based on the created input variables and customers are categorized as risky (future bad) or non-risky (future good) customers. Two different models are applied to predict the future performance. These models are as follows:

Model 1: A conventional stepwise logistic regression model is applied to demonstrate the significance of the created input variables in predicting future bad customers four months ahead.

Model 2: A hybrid SOM based classifier is proposed to predict the probability of future bad customers and the amount owed by these customers. The superiority of
the proposed classifier is verified by comparing the performance of the two models.

In this study we aim to show:

1. The usefulness of time series clustering based on a mixture of model-based and feature extraction methods in predicting customers’ future performance in behaviour scoring systems.

2. The usefulness of product identified expenditure in predicting customers’ future performance in behaviour scoring systems.

3. The superiority of a proposed Hybrid SOM based Classification Model as a behaviour scoring model over conventional (logistic regression) classification techniques.

1.6 Contribution

The contributions in this research relate mainly to the creation of input variables, the approach in selection of bad customers’ characteristics and a monthly view for prediction. Critical contributions in this research relate to the discovery of new variables for detecting delinquency, new procedures for predicting delinquency, and the use of only recent data for this purpose.

Creation of new input variables for detecting delinquency

The main contribution of this thesis is the creation of new input variables for credit scoring.

a) Product-base expenditure

We have created new variables which we believe have the potential to explain the financial circumstances of customers better than was previously the case. In
particular, product based expenditures are grouped to form input variables. We believe if a customer is not financially sound, he/she will not spend money on luxury products; instead his/her credit card will be used for purchasing only more necessary products such as petrol and food. These product-base variables represent the financial situation of customers.

b) Autoregressive parameters and moments

The transaction time series are not directly used in this study. A mixture of model based and feature based approaches are applied to daily transactional time series data allowing the extraction of parameters and moments from these time series.

**A new hybrid SOM-based classification approach for predicting delinquency**

A new approach for variable selection and model development is developed in this study. Customers are grouped into clusters based on transactional behaviour. The 3 clusters with the highest number of risky customers are determined and an appropriate model is designed to predict future performance of customers for each of these clusters. Each model captures the characteristics of risky customers within that particular group. The predicted probabilities of delinquency for these 3 clusters are added to the other input variables for the final model for the prediction of the probability of delinquency for any customer. A two stage model is used for the final model to predict the future performance (risky/non-risky) at the first stage and future balance owed at the second stage.

**One Month Observation Period**

There can be several time lags between the collection of transaction data and its use in a scorecard. In order to have enough history in a sample to decide if the customers are good or bad, a time period of six to twelve months is considered necessary in the banking industry. During this time, many changes could have occurred in the customer’s financial behaviour and in the economy. This is known as population drift which impacts the performance of current scorecard models.
Hence behaviour scoring systems need to be adjusted frequently to consider new situations.

The intention of this thesis is to develop an early warning system with only a month’s view of past behaviours. This fully utilizes all available transactional information collected in the previous month and predicts the future performance of customers using only this recent data. This time frame (one month view) uses only very recent data which identifies the present circumstances of the customer. Building a behaviour scoring model based only on the current financial circumstances of customers results in better prediction performance.

1.7 Outline of the thesis

In the next chapter, Chapter 2, the existing state of knowledge concerning credit and behaviour scoring systems is presented along with comparison studies, advantages and disadvantages of credit scoring methods.

Chapter 3 highlights the steps involved in developing a credit scoring system with a brief discussion of each step.

Chapter 4 provides a detailed literature review for the methodology used in this research.

Chapter 5 explains a pilot study performed on a small set of transactional data. Customers are classified into different groups based on transaction amounts, allowing the bank to distinguish between big spenders and small spenders. Also customers are grouped according to the types of products they purchase with their credit cards. The behaviour of the customers is analysed for 6 months and the movements from one cluster to another during this period are monitored, allowing the bank to understand changes in the financial circumstances of its customers from month to month.
Chapter 6 includes the data preparation for the main study. For this study transactional data, delinquency status and balances for approximately 1.5 million customers were considered. The detailed process used for the preparation of data and variables are explained in this chapter.

Chapter 7 includes the results for the main study performed on the data prepared as described in chapter 6. A hybrid SOM based classification model is developed to predict the future performance of customers and the future amount of money owed, so that very risky and less risky customers can be differentiated. This model is then applied to various populations in various contexts, allowing the approach to be evaluated in terms of its accuracy, robustness and versatility.

Chapter 8 is dedicated to the conclusion, a discussion of the limitations of the research and recommendations for future work. In particular, attention is paid to implementation issues and the benefits that banks can expect from this research.

The thesis also contains an appendix containing additional results and a SAS program to conduct the required analysis.
Chapter 2

LITERATURE REVIEW

IF THE FACTS DON'T FIT THE THEORY, CHANGE THE FACTS.

-THEORETICAL PHYSICIST ALBERT EINSTEIN (1879-1955)
Table 2.1  *Outline of chapter two*

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1  Introduction</td>
<td></td>
</tr>
<tr>
<td>2.2  Effect of credit cards on consumer credit</td>
<td></td>
</tr>
<tr>
<td>2.3  Overview of credit scoring</td>
<td></td>
</tr>
<tr>
<td>2.4  Behaviour scoring - The theory</td>
<td></td>
</tr>
<tr>
<td>2.5  Credit scoring systems in the past</td>
<td>Linear discriminant analysis (DA)</td>
</tr>
<tr>
<td>2.5  Credit scoring systems in the past</td>
<td>Regression</td>
</tr>
<tr>
<td>2.5  Credit scoring systems in the past</td>
<td>Logistic regression (LR)</td>
</tr>
<tr>
<td>2.5  Credit scoring systems in the past</td>
<td>Decision tree</td>
</tr>
<tr>
<td>2.5  Credit scoring systems in the past</td>
<td>Non-parametric models</td>
</tr>
<tr>
<td>2.5  Credit scoring systems in the past</td>
<td>Genetic algorithms (GA)</td>
</tr>
<tr>
<td>2.5  Credit scoring systems in the past</td>
<td>Artificial Neural Network (ANN)</td>
</tr>
<tr>
<td>2.6  Comparison of credit scoring methods and hybrid systems</td>
<td></td>
</tr>
<tr>
<td>2.7  Exploratory approaches for credit scoring data:</td>
<td>Markov chain (MC)</td>
</tr>
<tr>
<td>2.7  Exploratory approaches for credit scoring data:</td>
<td>Cluster analysis</td>
</tr>
<tr>
<td>2.8  Discussion</td>
<td></td>
</tr>
<tr>
<td>2.9  Limitations if credit scoring</td>
<td></td>
</tr>
</tbody>
</table>

**2.1 Introduction**

Even though credit scoring is one of the earliest financial risk management tools developed (Thomas, et al., 2002) and is commonly used in several industries, finding published literature on the method used is not easy. For reasons of competitive advantage and data confidentiality results are seldom published (Hand & Henley, 1997).

Most of the literature available on credit scoring is on application scoring, to the point that application scoring is automatically assumed when people talk about credit scoring.

Eisenbeis (1978) and Crook, Hamilton, and Thomas (1992) categorize the literature on credit scoring into two groups;
1. Literature on credit scoring which considers the merits of different techniques for constructing credit scoring models such as Linear discriminant analysis, Regression analysis, etc.

Some of the important contributions to the literature in this category include works by Durand (1941); Myers and Cordner (1957); Smith (1964); Myers and Forgy (1963); Altman (1968); Chatterjee and Barcun (1970); Orgler (1970, 1971); Edminster (1972); Apilado, Warner, and Dauten (1974); Wiginton (1980); Chandler and Coffman (1983/1984); Hand (1986, 2001); Leonard (1988, 1993); Srinivasan and Kim (1987); Boyle et. al (1992); Fogarty and Ireson (1993/1994); Hand and Henley (1993); Henley (1997); Chang et. al., (2000); Thomas (2000); and Hand and Kelley (2001). Relevant contributions from these authors are explained in the following sections.

2. Literature on credit scoring which considers the underlying objectives of credit scoring and the different phases of the credit granting process relating to credit scoring.

Among the most important contributions to the literature in this category are the works of Mehta (1968, 1970), Edelman (1992), Lundy (1992) and Oliver (1993). Relevant contributions from these authors are explained in the following sections.

Fogarty (2006) added a third class of literature on credit scoring which considers practical ways to improve the results of current scoring techniques. Recent work in this area includes that of Kelley, Hand, and Adams (1999), Hand and Adams (2000) and Fogarty (2006).

This thesis contributes to the third class of literature on credit scoring. We have used new variables, a different approach of variable selection and a combination of methods to obtain improved credit scoring results.

This chapter provides an overview of credit scoring methods used in practice. However, it is convenient to clarify again the difference between application scoring and behaviour scoring before we discuss the contents of the chapter.
Application scoring models (commonly called credit scoring models) are used to decide whether or not to grant credit to new applicants (Chen & Huang, 2003), whereas, behaviour scoring models (also known as maintenance or performance scoring models) are used to analyze the purchasing behaviour of existing customers (Setiono, Thong, & Yap, 1998). Both application and behaviour scoring deal with classification analysis and their main objective is to classify customers into groups consisting of people with similar default risk (Lancher, Coats, Shanker, & Fant, 1995). In credit scoring, classification analysis is applied to categorize a new applicant as “accept” or “reject” by using characteristics such as age, income and marital status (Chen & Huang, 2003), whereas classification of behaviour scoring is used to describe the behaviour of existing customers, based on behaviour characteristics such as payment patterns and spending patterns, and also to predict the future behaviour of existing customers (Setiono, et al.).

The standard techniques used in application scoring can be used for behaviour scoring. However, the data and the objective of behaviour scoring make it different from application scoring. This point is explained in detail in section 2.4.

The usefulness of credit scoring became particularly apparent with the advent of credit cards. In the next section, the effect of credit cards on consumer credit and credit scoring is discussed, followed by an overview of application scoring and behaviour scoring. Section 2.5 explains the statistical methods used to develop credit scoring systems in the past. Although there are other techniques used in credit scoring, in this chapter we discuss only techniques which are commonly used in practice. Several comparison studies are discussed under each credit scoring method. The techniques are explained in detail in chapter 4. The outline of the chapter is presented in Table 2.1.

### 2.2 Effect of credit cards on consumer credit

As said by Lewis (Lewis, 1992), consumer credit has been around since the time of the Babylonians, i.e., around 3000 years. The advent of credit cards in the 1950’s brought a new era for consumer credit. Credit cards made it possible for consumers to finance all their purchases at any time.
The concept of credit cards has evolved from credit schemes designed to sell fuel in the 1920s (mostly in the United States). A metal card was handed out by the petrol companies to customers who brought their goods. This metal card held the customer’s name, city and state but no address. In the 1950’s Diners Club, the first major credit card company, developed a card that could be used by any business that took credit cards. In the next two decades the credit card changed into a common payment method, allowing a “Cash Free” wallet which was advantageous for frequently traveling business people. Subsequently the increase in the number of credit card users and growth in credit card purchases was matched by the growth in credit card technology and hence credit cards became an essential part of life.

Credit cards are not only used because they allow borrowing on credit, but in today’s technology driven world credit cards makes life easier. Credit cards have become important money transmission methods. As mentioned previously the number of credit card users in Australia has increased from 6.5 million to 14.6 million in the last 15 years. In Australia the total value for credit transactions has increased from $160 million to $235 million per annum in the last 5 years. The growth in outstanding balances on credit cards has also been dramatic. In the last 25 years, the balance outstanding on credit card transaction has grown from $4.8 billion to $47.4 billion per annum. Figure 2.1 shows the outstanding balance on credit cards for the last 7 years in Australia.

![Credit Cards - Gross balance outstanding and growth (annual)](http://www.bankers.asn.au)

*Figure 2.1* Gross balance outstanding & annualized growth on credit cards
Increases in the number of credit cards, the value of transactions and the growth in outstanding balances has brought a new era for banks. Although banks benefit from this growth, they need to be careful about choosing the right customers. For banks, more customers and more outstanding balances indicate more risk. Consequently, choosing the right customers is the primary goal for banks. Many years ago, choosing the right customers was a judgmental decision based on credit references. The process of granting credit would take up to 9 days and the decision was as reliable as the perceptions and intuition of the bank manager. The introduction of credit scoring has changed this process and has improved its reliability.

2.3 Overview of credit scoring

Credit scoring is the term used to describe statistical methods used for classifying applicants for credit into good (non-risky) and bad (risky) classes. Such methods have become increasingly vital with the remarkable growth in consumer credit in recent years.

In the past decade, banks in many countries have been concerned about the process of upgrading their risk management performance and this has led to the development of methodologies, and the introduction of more careful practices, designed to measure and manage risk.

The history of credit scoring goes back to 1941, when Durand (1941) published his work on differentiating between good loans and bad loans based on results for 37 firms. Since then, credit scoring has become one of the most successful application areas for statistical and operational research.

As said in the previous section, in the past, decisions about granting credit were based on the personal judgments of credit analysts. During World War II, there was a shortage of credit analysts. Due to this shortage, the organizations made the analysts write down the rules they applied when granting credit to customers with good credibility in terms of repayment (Johnson, 2004). In particular, it was Henry Wells, manager of Alden, a
mail order house in Chicago, who pioneered credit scoring rules. He decided to have his credit scoring rules written down because he was losing his experienced credit staff. After the world war, organizations began to see the advantages of statistical models in decision making for purposes of granting loans. However, it was not until the creation of credit cards in the 1960’s that banks and other financial institutions recognized the advantages of credit scoring for decisions relating to the granting of credit. Organizations that started credit scoring realized that it was much more reliable than judgmental methods (Meyers & Forgy, 1963).

Credit scoring has many benefits over judgmental methods. It reduces discrimination since credit scoring uses variables which are only risk related. Variables such as race, religion and other discriminatory variables cannot be used in credit scoring models. Differentiating in the process of granting credit to customers without statistical justification was outlawed when the Equal Credit Opportunity Acts (1975; 1976) was passed in the USA. This served to validate the use of statistical models to provide a fair and impartial method for the granting of credit. Credit scoring avoids the use of personal subjectivity of credit analysts. Finally credit scoring automates and increases the speed and consistency of the credit granting process. With the success of credit scoring for credit cards, banks started to use credit scoring for other purposes such as personal loans, home loans and, as will be shortly discussed, behaviour scoring.

When credit scoring models were first introduced, the aim was to estimate future credit worthiness of applicants and to grant credit to those with low default risk. The underlying assumption for application scoring is that the credit worthiness of a customer is time independent (Thomas, 2000). Application scoring models are typically built using a minimum of one year’s credit performance of applicants. The data for application scoring is provided by credit bureaus. Nevertheless, in this competitive world, minimizing the default risk is only one of many concerns. It is also necessary to retain customers who are considered to be low risk customers and, while retaining these "quality" credit applicants, it is also important to maintain a profitable portfolio. In other words banks want to encourage good customers to spend more and pay more interest. This has led to the practice of behaviour scoring which is used to monitor the credit worthiness of existing customers.
2.4 Behaviour scoring - The theory

Credit has existed in various forms for many years but in recent times consumer credit has increased in the form of credit cards, home loans, personal loans etc. This has resulted in the widespread use of credit scoring. However, there are many aspects of the methodology that have not received sufficient attention in the research literature, due to the need for confidentiality resulting in a lack of availability of datasets for investigation purposes. In particular there is very little research literature for behaviour scoring. Both application and behaviour scoring rely on the development of classification tools using statistical analysis (Hand, 1981; Johnson & Wichern, 1998).

The standard statistical techniques used in application scoring can be applied to behaviour scoring models. However, the data and the objectives of behaviour scoring make it very different from application scoring. The objective of application scoring is to classify a new applicant as good (non-risky) or bad (risky) based on characteristics such as age, income, marital status, number of dependants and employment type, whereas behaviour scoring classifies the behaviour of existing customers based on purchasing and payment patterns. Behaviour scoring provides better information for setting credit limits, creating new products and identifying risky customers.

In this research we consider credit card portfolios and we aim to define a model which can describe the behaviour of credit card customers. In particular this model will allow us to identify customers whose behaviour is expected to change in the future.

The Basel Committee on Banking Supervision proposed a capital adequacy framework that allows banks to calculate capital requirements using “internal assessments” of key risk drivers (Angelini, di Tollo, & Roli, 2008). With Basel II the Basel Committee on Banking Supervision introduced a revised capital adequacy framework to support banks in making ongoing progress in their risk assessment capabilities rather than alternative regulatory standardized models (Basel Committee on Banking Supervision, BCBS, 2005). With the launch of Basel II more competition was brought to the banking world.
For the first time banks were allowed to rely on their own assessments of a borrower’s credit risk and therefore to employ the capital adequacy method most appropriate to the complexity of their transaction and risk profiles (Angelini, et al., 2008). Hence changes in the amount of capital held by banks are now directly reflected by the changes in credit quality. Therefore the need for more accurate systems to assess credit risk is even more important in today’s banking culture.

Large banks have developed sophisticated credit systems to assess credit risk arising from different portfolios and the search is still on. The more accurate risk measurement and better risk management contributes to more efficient capital allocation. As said by Ferguson (2001, p. 2):

“…..More accurate risk measurement and better management do not mean the absence of loss. Those outcomes in the tails of distributions that with some small probability can occur, on occasions do occur; but improved risk management has meant that lender and investors can more thoroughly understand their risk exposure and intentionally match it to their risk appetite, and they can more easily hedge unwanted risk…..”.

Banks need to maintain adequate capital as a cushion against any credit card losses. Banks therefore carry out their own internal risk assessments and determine the appropriate level of capital based on the credit-worthiness of their portfolio of credit card holders. The credit scores of existing and future customers are now used to establish what this level should be.

There are different methods for calculating the credit-worthiness of a borrower and accordingly the customer can be accepted or rejected for a loan or credit card. Even though the credit worthiness of the borrower is evaluated before the bank issues a credit card, there is still a lot of risk involved for the banks. Since banks need to maintain adequate capital as a cushion against any credit card losses, it is useful for banks to know and understand how a customer’s current behaviour can be used to forecast their future behaviour.
In regard to the credit card portfolio, when a credit card applicant is accepted a card is issued with a credit limit. The limit is set according to the apparent risk level of the applicant. Thereafter, the applicant’s transactions and payments are recorded in the bank’s database. Banks constantly monitor the transaction patterns of existing customers in order to understand the customer’s behaviour, allowing them to revise their evaluation for each customer’s credit worthiness over time. Fraudulent activity is also detected through this process (e.g., Credit card theft).

Banks are interested in understanding their customers’ spending and repayment patterns in order to help them minimize their losses. Behaviour scoring helps banks to estimate the probability that a customer’s credit behaviour remains in, or returns to, a satisfactory condition in the future. If a customer’s credit behaviour is not satisfactory, rehabilitation scoring can suggest whether it will be possible for the customer’s behaviour to revert to a satisfactory condition in the future. If banks know the answers to all these questions sooner rather than later, they can develop appropriate strategies for minimizing losses. Behaviour therefore makes use of a customer’s recent behaviour to predict whether or not they are likely to default in the immediate future. A pure behaviour scoring system will only include variables dealing with the customers’ performance and the current values of variables from monthly credit bureau reports. Other behavioural systems include personal characteristics such as age, time with the bank, residential status, as well as the pure behavioural characteristics.

According to Thomas, et al., (2004) a behaviour scoring model is developed using data for a sample of customers before and after a particular point in time, including all the characteristics which describe the performance of these customers over this period of time. The period before the observation time point is called the performance period and is usually 6 to 12 months. Typical variables included in the data include the average, maximum and minimum balance, credit turnover, and debit turnover. Some characteristics estimate the payment trend or balances during the performance period and some are indicators of delinquent behaviour – number of missed payments, times over overdraft or credit limit - while others reflect difficulty in money management such as the number of cash advances using a credit card.
The period after the observation time point is the outcome period, which is usually taken as 12 months, and the customer is classified as a good or a bad customer depending on their status at the end of this outcome period. A common definition is to classify as bad a customer who is 90 days overdue at this point. In order to separate the good and bad customers as much as possible for behaviour scoring purposes, those with behaviour that is not yet bad but is tending that way are classified as indeterminate. Thus customers who are 30 to 90 days overdue at the end of the outcome period may be excluded for behaviour scoring purposes so that the sample contains only good customers who are less than 30 days overdue and bad customers who are 90 or more days overdue at the end of the outcome period.

As described by Thomas, et al., (2004) above, behaviour scoring models are typically built using a two year history for customers. However, this may result in a performance period which is very different to the outcome period. He recommends shortening the performance period to 6 months in order to overcome this problem.

For existing customers, behaviour scoring helps banks make well-informed decisions using forecasts of future customer performance. Behaviour scores are not only used to identify the risky customers, they are also used in assigning new credit limits for good customers, marketing new products to good customers, or managing recovery of debt if an account turns bad.

In the banking industry time plays a crucial part. It is important to understand the customer’s behaviour, but it is equally important to predict the changes in a customer's behaviour in a timely manner. If banks have advance warning about changes in the financial behaviour of customers, they have adequate time to come up with intelligent tactics to deal with these changes.

This research aims to identify the change in financial behaviour of customers as early as possible.

Statistical methods such as multiple regression, linear discriminant analysis and decision trees are still the most popular methods for credit scoring. However, in the last
decade, expert systems and neural networks have also been of great interest for credit scoring models. Given today’s competitive environment and with the rapid growth of information processing technology, increasing complexity of financial market activity and the dramatic growth of credit users, banks are not satisfied with only applying the popular methods. There is a need for more sophisticated strategies and approaches which not only involve choosing the right (low risk) applicants but also suggest tactics to minimize losses as quickly as possible and to attract and retain customers whilst maintaining a profitable portfolio (Siddiqi, 2006).

Hence the search is still on for better models that not only estimate the creditworthiness of applicants and existing customers but also take into account bank profitability and other concerns.

### 2.5 Credit scoring systems in the past

Credit scoring has been used for many years. Before and during World War II various attempts were made to have credit decision making tools (Lewis, 1992), but it is only in the last 20 years that well designed techniques for credit scoring have become available, and these techniques have only become commonly used in the last 10 years. The most widely used techniques for building score cards are linear discriminant analysis and linear regression since both of these methods are theoretically straightforward and also they are widely available in statistical software packages. Other techniques which have been used in the industry include logistic regression, probit analysis, nonparametric methods, mathematical programming, Markov chain models, recursive partitioning, expert systems, genetic algorithms, neural networks and conditional independence models.

In the 1950’s, score card systems were developed for credit scoring, but the choice of variables and weightings for scoring purposes was judgmental or guided by rules of thumb (Johnson, 1992). As said previously the usefulness of credit scoring was brought to the fore by the advent of credit cards in the late sixties when the burgeoning number
of people applying for credit cards left the banks no choice but to automate the lending decision. In the late 1960's and early 1970’s, economic pressure and computer technology forced the acceptance of empirical valuation methods for credit scoring (Thomas, et al., 2002). A study by Chandler and Coffman (1979) highlighted certain advantages of credit scoring over judgmental methods. This study suggested that these empirical methods are based on actual performance and can therefore be statistically validated before implementation. Also this study found that these empirical methods, on average, are more accurate compared to judgmental methods. The agencies who used credit scoring found it a better predictor of default than judgmental decisions and this led to default rates dropping by up to 50% (Meyers & Forgy, 1963).

The credit scoring literature is discussed below in relation to seven commonly used approaches for predicting delinquency namely, linear discriminant analysis, regression, logistic regression, decision trees, non-parametric models, genetic algorithms and artificial neural networks.

*Linear discriminant analysis (DA)*

One of the most common methods for building credit scoring systems is Linear discriminant analysis (LDA). The theory of discriminating between different groups in populations, based only on a few related characteristics, was introduced by Fisher (1936) with the famous iris experiment. In this experiment Fisher found that the physical size of plants could be used to differentiate between three varieties of iris. David Durand (1941) was the first to recognize that this technique could be used to identify good and bad loans. It was then Wonderlic in 1948, who realized that, if scored properly, discriminant scores could be used to evaluate the risk for personal loans (Wonderlic, 1952).

LDA has the advantage of being simple conceptually and available in most statistical software packages. However, one of the drawbacks in linear discriminant analysis is that it does not deal well with dependencies between the variables, and it uses all the selected variables even when some of these variables are redundant (Boyle, et al., 1992). Another significant limitation of LDA is that it assumes that the predictor
variables are normally distributed and have linear and homoscedastic relationships. This means that LDA cannot handle independent characteristics which are of a categorical nature. In credit scoring, the aim is to identify the ‘credit–worthy’ and ‘non-credit-worthy’ customers. The covariance matrices of these two classes are not likely to be equal, and in many cases the characteristics are categorical in nature, hence the use of LDA is often not appropriate (West, 2000). However, in a comparison study by Yobas, Crook, and Ross (1997) it was found that the predictive performance of LDA was superior to many other machine learning methods such as artificial neural network, generic algorithms and decision trees. Many studies have highlighted the problems when using DA for credit scoring. The seven statistical problems in applying DA in credit scoring raised by Eisenbeis (1977, 1978) are summarized below.

i. Non-normal Distribution for the variables:

DA assumes that the variables are multivariate normally distributed. In the applied literature, this assumption has been ignored, mainly due to the lack of a test for multivariate normality. However, there are studies such as Myers and Forgy (1963) which argue that DA is robust to small departures from normality. More so, the variables used in credit scoring are often categorical in nature making it necessary to recode them into binary variables. Violation of the normality assumption may bias tests of significance and the estimated error rates.

Now the problem is how to treat the variables once we identify departures from normality. A common approach is to use an appropriate transformation before estimating the Discriminant function (Pinches & Mingo, 1973). However Eisenbeis (1977) points out, that applying transformations may change the interrelationships among variables and maybe affect the relative positions of the observations in the group. In addition it is not possible to find a transformation which can produce a normal distribution for a binary variable.

4 In a study by Reichert, Cho & Wagner (1983), the authors argue that, from the pragmatic point of view, the non-normality of credit information is not a limitation for the use of LDA
ii. Unequal covariance matrices for groups in population:

Linear discriminant analysis assumes equal covariance matrices across all groups. Violation of this assumption affects the significance test for the differences in group means and it also affects the appropriate form of classification rules. When the covariance matrices of the two groups are unequal, Quadratic Linear discriminant analysis should be used instead of Linear discriminant analysis. Therefore it is important to check the equality of covariance matrices before constructing the classification model. In the credit scoring literature it seems that this is rarely done (Henley, 1995). However, it should be noted that quadratic rules are more sensitive to deviations from normality than linear rules. Furthermore, Monte Carlo studies by Marks and Dunn (1974) show that, when samples are small and relatively large number of variables are involved, Linear discriminant analysis performs better than Quadratic Linear discriminant analysis even if population covariance matrices are unequal.

iii. The significance of variables in the classification model:

There is a problem with determining the relative importance of individual variables in DA. There is often a need to demonstrate which variables contribute significantly to the discriminatory power of the model. This can be done for a particular variable by using a significance test to compare the performance of the model with all the variables included in the model and then omitting the variable of interest. However, this approach has not been automated in most statistical software packages.

iv. Group definitions:

DA assumes that the groups are discrete and identifiable. In credit scoring systems, commonly the customers are classified as “good” risk or “bad” risk, which satisfies this necessary condition for using DA, because there are distinct and non-overlapping group

---

5 Covariance matrix: a covariance matrix is a matrix whose element in the \(i,j\) position is the covariance between the \(i^{th}\) and \(j^{th}\) elements of a vector of random variables. Covariance is a measure of how much two variables change together.
definitions. But this is not always the case (e.g., refer to Eisenbeis, (1977)) for further details).

v. Use of inappropriate a priori probabilities and costs of misclassification:

For performing linear discriminant analysis a prior probabilities of an applicant belonging to the good and bad classes and also the relative costs of misclassification for the two classes must be specified. Little emphasis is given to this rule and very often a priori probabilities for the groups and the costs of misclassification for good and bad customers are assumed to be equal. This causes the misclassification rate to be a poor estimate of the true error rate. Use of appropriate a priori probabilities is critical for the evaluation of the performance of the model and for default prediction.

vi. Estimation of classification error rates:

The main concern in using DA as a classification method is assessing the performance of the estimated rule. In the literature, especially in finance and economics, it is publicized that biased results are obtained if the reclassification approach is used to estimate the expected error rate. There are a number of alternative methods. Hand (1986) and Toussant (1974) review different approaches for estimating the error rate. The most common and most expensive approach is the hold-out-method. A hold-out method randomly divides the available sample into two parts; an analysis sample, used to perform the linear discriminant analysis, and a hold-out sample (also called a test sample) used to estimate the error rate. The majority of credit scoring systems appear to use this hold-out method.

Reichert, Cho, and Wagner (1983) demonstrate that using the holdout procedure to choose the best classification rule leads to reasonably consistent and reliable discriminant coefficients. However using the reclassification approach for estimating

---

6Reclassification approach is when the original analysis sample is used to estimate the expected error rate.
error rates leads to a bias of about 2-3 % compared to using a 50% random holdout procedure.

vii. Selection of analysis samples:

One of the main problems in building credit scoring models is the selection of representative analysis samples. The sample used to construct the scoring system often only consists of accepted applicants since the true creditworthiness of the rejected applicants is usually unknown. This results in a biased scorecard. It is important that “bad” cases should represent at least 20% of the sample according to Berry and Linoff (2000).

Although the use of linear discriminant analysis for credit scoring has often been criticized due to the categorical nature of credit data and unequal covariance matrices (West, 2000), some authors have found that the predictive performance of DA in credit scoring is superior to that of artificial neural networks, genetic algorithms and decision trees (Yobas, et al., 1997), provided that the other issues raised above have been addressed.

It is important to note that, in the generation of predictors, the distribution theory is not crucial since the proof of choice of predictors is evaluated by their use. Therefore the fact that LDA or other methods such as linear regression analysis assume normality is not very important.

Another important point to highlight is that prior probabilities are a requirement for any decision making framework. However appropriate prior probabilities are mostly not known. Therefore the disadvantage of inappropriate priors and costs of misclassification are not only a disadvantage of DA, they are a disadvantage of all classification methods.
Regression

Ordinary Least Square (OLS) Regression is another technique used in credit scoring systems. Myers and Forgy (1963) compared linear discriminant analysis and stepwise regression. They assigned equal weights to each of the predictive characteristics selected by the stepwise regression and concluded that there could be ways to improve the effectiveness of the linear discriminant analysis approach. Orgler (1970) also created a credit scoring model using a linear regression, but this was for commercial loans.

The linear regression technique uses the data very efficiently. With relatively small datasets one can obtain good results. Also Linear Regression is well-understood and easily interpretable. OLS regression is not suitable for estimating the probability of default because the estimates will not be bounded by 0 and 1 and also, in general, it is not efficient for extrapolation outside the range of the existing data. However, it can be used to predict the magnitude of account balances.

Logistic Regression (LR)

In the 1980’s, Logistic Regression (LR) was introduced for credit scoring. Wiginton (1980) was the first to apply logistic regression for credit scoring, and, in comparison with linear discriminant analysis he concluded that logistic regression gave superior results. Wiginton (1980) explained that logistic regression is more appropriate for credit scoring since, while it allows the assumption that creditworthiness (dependent variable) has a Bernoulli distribution, it does not assume that the characteristics of creditworthiness are drawn from a multivariate normal distribution. In his study on data collected by an oil company, he compared the performance of a logit model and Linear discriminant analysis. The logit model achieved a 62% correct classification compared with an expected 58% correct classification using chance (i.e., allocating all cases to the largest group). Linear discriminant analysis gave the same performance as a classifier based upon chance. However, in his study, the number of variables was unrealistically small. Henley (1995) found that logistic regression was little better than LDA for his credit scoring data because a large proportion of the applicants in his data had scores
between 0.2 and 0.8 as the estimated probabilities for a good risk customer. In such a case the logistic curve is approximated by a straight line. A comparison of probit analysis (logistic regression with a probit link function) and lineardiscriminant analysis was performed by Grablowsky and Talley (1981). The comparison was conducted on a dataset from a retail chain in the U.S. with eleven explanatory variables which were found to be predictive of delinquency over a ten year period. Four of these variables were included in the model using Stepwise linear discriminant analysis. A likelihood ratio test was performed to test the individual variable’s significance in probit analysis which resulted in the selection of nine significant variables for the model. Table 2.2 shows the classification rate for their study.
Table 2.2  
*Classification results from a study by Gravlowsky and Talley (1981)*

<table>
<thead>
<tr>
<th>Actual class</th>
<th>Probability for Predicted class using Linear discriminant analysis</th>
<th>Probability for Predicted class using Probit analysis</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good 0.10 0.00</td>
<td>Good 0.91 0.09</td>
</tr>
<tr>
<td>Good</td>
<td>Bad 0.20 0.80</td>
<td>Bad 0.07 0.93</td>
</tr>
<tr>
<td></td>
<td>Total 1.00</td>
<td>Total 1.00</td>
</tr>
</tbody>
</table>

Clearly probit analysis achieved better performance compared to linear discriminant analysis provided that misclassification of the bad risk individuals is more costly than misclassification of the good risk individuals. Logistic Regression was also used successfully for the granting of corporate credit (Srinivasan & Kim, 1987) and in commercial loan evaluation (Leonard, 1993).

**Decision tree**

The Recursive Partitioning Algorithm (RPA) – also known as a decision tree analysis has been used extensively in credit scoring and has resulted in good performance in many studies.

Tree applications in credit scoring are described by Mehta (1968), Makowski (1985), Carter and Catlett (1987), and Davis, Edelman, and Gammerman (1992).

A general comparison study of several scoring systems was carried out by Srinivasan and Kim (1987). LDA, LR, Goal Programming (GP) and RPA were used. This study was conducted on commercial loans and hence the dataset was small (215 cases with only 8 variables). Bootstrapping was used to estimate the error rate. This technique is commonly used when the data violates distributional assumptions. The results showed that RPA performed best and the logistic model gave marginally better performance compared to GP and LDA.

A similar study was performed on consumer credit data by Boyle et al. (1992). The statistical methods used for this comparison study were DA, RPA and hybrid methods using both techniques. The dataset consisted of 1001 accounts and 24 variables (application characteristics) and the aim was to identify slow payers amongst existing credit card holders. The hybrid classifier involved constructing two decision trees to
identify interactions between the original characteristics which were then included in a linear discriminant model as characteristics. This classifier performed best on the holdout sample and the decision tree characteristics were found to be the most important characteristics in the linear discriminant analysis. It was also found that LDA performed marginally better that RPA for the test data. In particular, DA is a successful classifier since it uses the whole data in building the model and the strength of RPA is that it permits the modeling of complex dependencies between variables. The author commented that both DA and RPA have complementary strengths and can be combined to give a successful hybrid classifier.

Finch and Schneider (2006) compared the classification accuracy of LDA, QDA, Logistic regression and tree classification using Monte Carlo simulations. They reported that the misclassification rate using the QDA approach was never larger than LDA and LR and it was lower in many cases. When the data was normally distributed and the covariance matrices of the groups were equal, LDA, LR and QDA had similar misclassification rates. However, they recognized that tree results showed higher error rates than the other three models. Also in their study they observed that if the data conditions were not met, the error rates for LDA and LR went up, and when the covariance matrices were not equal QDA and tree misclassification rates went down (Finch & Schneider).

*Non-parametric models*

A study by Chatterjee and Barcun (1970) is one of the early examples of nonparametric approaches to credit scoring. The study was performed on personal loan applicants of a bank in New York. The aim was to classify an applicant to the class with which he/she had most in common, so as to minimize the expected loss from misclassification. The classification error rate was estimated using the jackknife method.

From nonparametric methods, the nearest neighbor method is the most common method used in credit scoring (Chatterjee & Barcun, 1970; Hand, 1986; Henley & Hand, 1996). Henley and Hand (1996) estimated the probability of default using a k-nearest
neighbour estimator. The target value for the majority of the k-nearest neighbours for any customer determines the good/bad prediction. In their study, they concluded the superior performance of this estimator over logistic regression, linear regression and a decision tree. Also they argued that nonparametric estimation is able to model the irregularities in the risk function over the feature space and that parametric models fail to capture these irregularities because of their distribution assumptions.

One of the advantages of the nearest neighbor method for credit scoring is that it is straightforward to update directly by adding new customers to the set when their true class is known and deleting old customers (Henley & Hand, 1996). This helps to overcome the problems with changes in population over time. In other words, it can overcome the population drift7 which is a major problem in building credit scoring models. On the other hand, it requires a major system investment due to very intensive computational operation of searching for the k closest neighbors.

**Genetic Algorithms (GA)**

The advances in computing and need for sophisticated scorecards, have led to the implementation of intelligent techniques in credit scoring. It is interesting how biology has become a great source of inspiration for computational problems. Over the years, many intelligent systems have been constructed based on biological metaphors. These methods include genetic algorithms and neural networks. Generic Algorithms (GA) are complicated computing methods based upon a biological metaphor of evolution. GA was first proposed by John Holland in 1960s. For more information on GA refer to Holland (1974) and South, Wetherill, and Tham (1993).

“Genetic Algorithms (GAs) are heuristic randomised search algorithms reflecting the survival of the fittest principle that can be observed throughout many evolution processes in nature” (Schlottmann & Seese, 2001, p. 7).

---

7 Population drift is the changes to the customer’s behaviour over time which could be due to economic pressure or marketing changes. This issue is discussed in section
Fogarty and Ireson (1993/1994) have shown the potential of the generic algorithm (GA) in building credit scoring models. GAs were thought to be appropriate for credit scoring since they can efficiently search large solution spaces when conventional statistical theory is inappropriate. The authors compared different classifiers on a large dataset including 51,020 applicants for consumer credit with 94% good and 6% bad applicants. This dataset was randomly split into ten equal test sets. At each stage, a test set was selected and classifiers were constructed on the remaining datasets. The classifiers were used to estimate performance measures on the test set and then the results were averaged over the test sets. Classification accuracy – proportion of good credit customers and bad credit customers correctly classified, profitability and acceptance rate were used as measures of performance. A profitability criterion was calculated using the number of good credit customers correctly classified minus 8 times the number of bad credit customers incorrectly classified, all divided by the total number of good credit customers. This criterion gives a weighting of 8 to the bad applicants and therefore removes the advantage of the default rule. According to the profitability results, the genetic algorithm outperforms the other classifiers. An acceptance rate is considered to explain the differences in classifier ranking between the classification accuracy and profitability results. The proportion of bad applicants accepted will decrease rapidly as acceptance rate decreases. Therefore classifiers with relatively low acceptance rates are likely to have a relatively higher profitability rates than classification accuracy rates since the profitability criterion gives priority to bad applicants. In this study, it appears that GA performs well giving the best profitability and the closest accuracy to the default rule.

A study by Yobas, et al. (1997) compared the predictive performances of four techniques, in identifying good and bad credit card holders. The techniques under study were ANN, GA, LDA and decision trees. Interestingly, in this study, it was found that linear discriminant analysis performed best, followed by GA.

Other publications on the application of GAs in credit scoring include those of Chen & Huang (2003) and Varetto (1998). South et al., (1993) highlights the reasons why genetic algorithms are useful in classification problems such as credit scoring:
1. Genetic algorithms reduce the chance of converging to local optima. This is because GAs do not search for a single solution point, they rather search from a population of solutions and they are able to search new parts of the solution space by using crossover operator.

2. GA uses probabilistic rather than deterministic transition rules for generating new solution sets. Probabilistic models consider randomness and variable states are not described by unique values, but rather by probability distributions which is more suitable for credit data.

3. GAs work on a set of potential solutions and not on a single solution.

4. GAs use available scoring information (objective function information) rather than derivatives or other auxiliary knowledge to assess the value of potential solutions.

Golberg (1989) discusses the effectiveness of GAs in searching large datasets. GAs learn complex relationships in incomplete datasets and can discover unknown patterns (Goonatilake & Treleaven, 1995) and hence they are suitable for credit scoring. GAs have been used as scorecards in a credit union environment (Desai, Conway, Crook, & Overstreet, 1997).

With genetic algorithms there is a big possibility of over fitting a model. There is no absolute assurance that a genetic algorithm will find a global optimum. Calculation is complex and this method results in lower discriminatory power compared to discriminant analysis and neural network (Thomas, 2002; Anderson, 2007).

Because of these disadvantages and the fact that GA is rarely used in credit scoring we have not explored this method in this research. We have focused on more commonly used methods that are readily available to banks and financial institutions.

**Artificial Neural Network (ANN)**

Artificial Neural Networks (ANNs) or Neural Networks (NN) are inspired by the structure and functionality of the brain nerve cells. The history of neural networks goes back to the 1940’s. The first model of the biological neuron was proposed by the neurophysiologist Warren McCulloch and the logician Walter Pitts in 1943 (McCulloch
& Pitts, 1943). This model was then utilized in the development of the first artificial neural network by Rosenblatt in 1959 (Rosenblatt, 1959; Widrow & Hoff, 1960) which was based on a unit called the perceptron. Variations on the perceptron-based artificial neural network were further discovered in 1960s (Rosenblatt, 1962; Widrow & Hoff, 1960). Interest in ANN disappeared discharged in the 1970’s with the untimely death of Rosenblatt and the demonstration of Minsky and Papert that the perceptron was unable to represent simple linearly inseparable functions (Minsky & Papert, 1969). However, the interest in ANNs was refreshed in the 1980’s with the realization that adding (hidden) layers to the network could yield significant computational versatility (Schalkoff, 1997). With the advances in computer technology and improvements in our understanding about brain functions, the application of ANNs increased. The exploration of neural network has taken on a significantly pragmatic feel in recent years such that there is greater interest in using neural networks as problem-solving algorithms than in developing them as accurate symbols of the human nervous system (Ripley, 1994).

The main advantage of ANNs is that, with a suitable number of hidden nodes they can approximate nonlinear functions to an arbitrary degree of accuracy (Hornik, Stinchcombe, & White, 1989) and provide reliable ways of obtaining solutions to a variety of problems that often cannot be dealt with using traditional methods (Goonatilake & Treleaven, 1995). NN is a data-driven approach and therefore pre-specification of the model is not required. This is where NN gain advantage over traditional approaches in credit scoring. As said by Allen Jost (1993, p.32):

“Traditional statistical model development includes time consuming manual data review activities such as searching for non-linear relationships and detecting interactions among predictor variables”.

At present, much research on neural networks is taking place within two areas; the multilayered feed-forward networks (also known as multilayer perceptrons) which is used for classification problems and symmetric recurrent networks (also known as Hopfield nets) which are used for developing associative memory systems (Ripley, 1994). ANNs are successfully applied in finance and banking. ANN has recently been
claimed to be an accurate tool for credit analysis in the credit industry among others (Desai, Crook, & Overstreet, 1996; Malhotra & Malhotra, 2003; West, 2000).

Many studies have used neural networks in credit scoring and the superiority of this technique in comparison with Linear discriminant analysis, logistic regression analysis and other traditional statistical methods (Chuang & Lin, 2009; Coats & Fant, 1993; Desai, et al., 1996; Huang, Chen, Hsu, Chen, & Wu, 2004; Jensen, 1992; Lancher, et al., 1995; Malhotra & Malhotra, 2003; Sharda & Wilson, 1996) has been established. Rosenberg and Gleit (1994) described several applications of neural networks to corporate credit decisions. Desai, Crook and Overstreet (1996) investigate the performance of a multilayer perceptron neural network, a mixture of an expert’s neural network, logistic regression linear and linear discriminant analysis for credit scoring in the credit union industry. Their results indicate that, if the measure of performance is the percentage of bad loans correctly classified, then customized neural networks offer a very promising opportunity. If the measure of performance is the percentage of good and bad loans correctly classified, logistic regression models are comparable to the neural networks approach.

Chen and Huang (2003) compared neural networks, LDA and classification trees. The classification results showed that NNs perform better compared to the other two methods if the measure of performance is the percentage of good and bad credits correctly classified. They also mention that if the measure of performance is percentage of good credit customers accurately classified, LDA is best followed by classification tree. If the measure of performance is percentage of bad credits accurately classified then BP performs better than LDA and a classification tree.

2.6 Comparison of Methods for Credit Scoring and Hybrid Systems:

Table 2.3 provides a summary of the studies previously conducted which compared the efficacy of these methods. Different data sets were used in these studies, justifying the variation found for the preferred method.
<table>
<thead>
<tr>
<th>Author</th>
<th>Year</th>
<th>Comparison Methods</th>
<th>Preferred Method</th>
</tr>
</thead>
<tbody>
<tr>
<td>Myer and Forgy</td>
<td>1963</td>
<td>DA, LinR</td>
<td>LinR</td>
</tr>
<tr>
<td>Wiginton</td>
<td>1980</td>
<td>DA, LR</td>
<td>LR</td>
</tr>
<tr>
<td>Henley</td>
<td>1995</td>
<td>LinR, LR, RPA</td>
<td>RPA</td>
</tr>
<tr>
<td>Boyle et al.</td>
<td>1992</td>
<td>LinR, RPA, LP</td>
<td>LinR</td>
</tr>
<tr>
<td>Srinivisan &amp; Kim</td>
<td>1987</td>
<td>LinR, LR, RPA, LP</td>
<td>RPA</td>
</tr>
<tr>
<td>Yobas et al.</td>
<td>1997</td>
<td>LinR, RPA, NN, GA</td>
<td>LinR</td>
</tr>
<tr>
<td>Desai et al.</td>
<td>1997</td>
<td>LinR, LR, RPA, NN</td>
<td>LR</td>
</tr>
<tr>
<td>West</td>
<td>2000</td>
<td>LDA, LR, RPA, NN</td>
<td>NN</td>
</tr>
<tr>
<td>Lee et al.</td>
<td>2002</td>
<td>LDA, LR, NN, NN(^a)</td>
<td>NN*</td>
</tr>
<tr>
<td>Malhotra &amp; Malhotra</td>
<td>2003</td>
<td>LDA, NN</td>
<td>NN</td>
</tr>
<tr>
<td>Baesens</td>
<td>2003</td>
<td>LDA, LR, RPA, NN, Knn, SVM(^b)</td>
<td>NN</td>
</tr>
<tr>
<td>Ong et al.</td>
<td>2005</td>
<td>LDA, RPA, NN, GP</td>
<td>GP</td>
</tr>
<tr>
<td>Chen and Huang</td>
<td>2003</td>
<td>LR, RPA, NN</td>
<td>NN</td>
</tr>
<tr>
<td>Desai et al.</td>
<td>1996</td>
<td>LDA, LR, NN</td>
<td>NN</td>
</tr>
<tr>
<td>Chuang and Lin</td>
<td>2009</td>
<td>DA, LR, RPA, NN, NN(^c)</td>
<td>NN(^c) followed by NN</td>
</tr>
<tr>
<td>Ince et al.</td>
<td>2009</td>
<td>LDA, LR, RPA, NN</td>
<td>RPA</td>
</tr>
</tbody>
</table>

LinR: Linear Regression  
LR: Logistic Regression  
RPA: Recursive partitioning algorithm (Classification Tree)  
LDA: Linear discriminant analysis  
NN\(^a\): Hybrid LDA and NN.  
NN\(^b\): two stage reassigning credit scoring model involving Multivariate adaptive regression splines to select significant variables, then NN is used to classify customers as good and bad. At second stage the rejected applicants are reassigned to the conditional accepted class (see- Chuang & Lin 2009).  
SVM\(^b\): Support vector machines are a set of related supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. This method is not discussed in this thesis.  

Studies by Hand and Henley (1997) and Henley (1995) are the most detailed published analyses in which various methods of credit scoring such as DA, LR, mathematical programming and smoothing non-parametric methods are compared. In their study it was concluded that there is no overall best method for credit scoring, unless the type of the problem is known. For example, NN are best for situations where there is poor
understanding of the data structure, whereas regression and tree classifications are more appropriate in situations where reasons are necessary for any decisions. West (2000), examined the accuracy of different quantitative models in credit scoring. In his investigation, he found that NNs improve the accuracy of credit scoring by up to 3%. Also he reported that logistic regression is a good alternative to NNs.

Ince and Aktan (2009) looked at the performance of credit scoring using linear discriminant analysis, logistic regression, neural networks and classification and regression trees (CART), performed on a credit card data set. Results reveal that CART has a better average correct classification rate in comparison with the other three methods. Table 2.4 shows the predicted accuracy of the models.

<table>
<thead>
<tr>
<th>Models</th>
<th>DA</th>
<th>LR</th>
<th>NN</th>
<th>CART</th>
</tr>
</thead>
<tbody>
<tr>
<td>Predicted Accuracy</td>
<td>62.20%</td>
<td>62.33%</td>
<td>61.52%</td>
<td>65.58%</td>
</tr>
</tbody>
</table>

On the other hand, if cost of misclassification is considered neural network credit scoring model has better overall credit scoring capabilities. This is because the cost of misclassifying bad customers as good customers (Type II) is higher than misclassifying good customers as bad customers (Type I) (Desai, et al., 1996). Since neural network credit scoring model has lower customer with bad credit misclassified as good customers, this model is selected as best compared to the other three methods. Table 2.5 shows the probability of Type I and Type II errors for the models under study.
Table 2.5 *Probability of Type I and Type II Errors of Different Credit Scoring Models in a Comparison Study by Ince et al. (2009)*

<table>
<thead>
<tr>
<th>Credit scoring models</th>
<th>Type I error</th>
<th>Type II error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Linear discriminant analysis</td>
<td>31.90%</td>
<td>43.32%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>42.86%</td>
<td>32.22%</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>44.59%</td>
<td><strong>29.25%</strong></td>
</tr>
<tr>
<td>CART</td>
<td>33.01%</td>
<td>39.88%</td>
</tr>
</tbody>
</table>

Another comparison study was conducted by Ibrahim and Wah (2010) to explore the performance of credit scoring models using Logistic Regression, Decision Tree and Neural Network to classify credit card applicants. Comparisons of the performance of the credit scoring models were based on misclassification rate. In their study results revealed that Decision Tree using chi-square splitting criteria performs best since it has the lowest misclassification rate (20.18%). Table 2.6 shows the misclassification rate of the models in their study. They also advise that the performance of predictive models depends on the data structure, data quality and the objective of the classification. The authors mention that in practical applications, classification methods such as decision trees and logistic regression are more appealing to users since they are comparatively easy to understand and deploy; however with the accessibility of data mining softwares, banks are finding data mining techniques useful in credit scoring.

Table 2.6 *Misclassification Rate of Different Credit Scoring Models by Irma & Yap (2010)*

<table>
<thead>
<tr>
<th>Models</th>
<th>Misclassification Rate for Validation data</th>
<th>Misclassification Rate for Training data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tree</td>
<td><strong>20.19 %</strong></td>
<td>19.39 %</td>
</tr>
<tr>
<td>Neural Network</td>
<td>24.29 %</td>
<td>19.82 %</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>28.07 %</td>
<td>27.16 %</td>
</tr>
</tbody>
</table>

Reports from rating agencies such as Moody’s (Khandani, Kocagil, & Carty, 2001) and Standard & Poor’s (Friedman, 2002) suggested that although NNs are successfully
implemented in the field of credit scoring techniques, their application is limited because NNs are not easily interpretable and they need a large training set.

The above literature review refers to credit scoring methods found in the literature. The next section refers to two exploratory methods for dealing with credit scoring data, namely Markov Chain, Cluster Analysis.

2.7 Exploratory Approaches for Credit Scoring Data:

Markov Chain

A Markov chain (MC) has an advantage in that it gives a probability of belonging to a range of states after a fixed period of time, rather than a simple good or bad prediction. Also Markov chains take into account the multi-period problem of credit granting—a point raised by Eisenbeis (1978). The decision to grant credit is not a one-period problem. It can affect the value of the customer relationship over future time periods, creating or preventing a multi-period problem.

A study by Cyert, Davidson, and Thompson (1962) is one of the earliest studies using stationary Markov chains in credit scoring based on repayment behaviour. In their study they assumed that in each month an account can belong to one of (n+2) states (paid, current, one month overdue,…,(n-1) month overdue, bad debt). A transition probability matrix is defined to show the probability that an account in one state moves to another state. They used maximum likelihood estimation to calculate the elements of thus matrix. Cyert and Thompson (1968) used separate transition matrices for different risk classes and credit limits are calculated over n period to maximize a profit criterion. For more work on application of stationary Markov chains the reader can refer to Corcoran (1978) and van Kuelen et al. (1981).

Cluster analysis

Supervised learning methods require predefined target classes. The advantage of supervised learning methods is that they can achieve high prediction accuracy if the
historical data and the data under study are similar in characteristics. However, the
disadvantage of supervised learning methods is that they can only be used only if the
data label for the target class is known. When the data label is unknown, unsupervised
learning methods could be useful. Clustering is an unsupervised learning method,
which groups the data according to the underlying characteristics and structure.

An early study using clustering in credit scoring was by Lundy (1992). The main
purpose of this study was to identify subgroups of the population for marketing
purposes rather than credit scoring. A second application of clustering to credit granting
is a study by Edelman (1992). In our study we use clustering for identifying the clusters
of clients with a high probability of defaulting so that the behaviour of risky customers
within these clusters is separately identified. More recently, Peng, Kou, and Shi (2005)
used clustering analysis and compared the results with supervised classification
methods. They concluded that cluster analysis as a stand-alone analysis, results in lower
classification accuracy than supervised learning methods. On the other hand, they
concluded that if clustering is combined with supervised methods, the classification
results can be improved significantly. Our study has confirmed this conclusion.

2.8 Discussion

Credit scoring has become a major facet of today’s society. The significant increase of
credit scoring is due to the expansion of consumer credit in the past few decades. Due to
these automated processes of decision lending, credit decisions are made much quicker,
which results in cost saving for banks.

The available literature on credit scoring generally deals with a description of the
techniques used to calculate the probability of default and a comparison of the
techniques in regard to misclassification rates. Descriptions of the main techniques used
in credit scoring can be found in Hand and Henley (1997), Rosenberg and Gleit (1994)
and Reichert et al. (1983).
Traditionally, LDA and Linear Regression are the most widely used techniques in building credit scoring models. However, because of the categorical nature of the data used for credit scoring and the fact that the covariance matrices of the groups (good credit customers/bad credit customers) are not likely to be equal, the use of LDA is been criticized for credit scoring. Logistic Regression is by far, the most common applied method in building a scorecard in today’s world according to many scorecard analysts. Wiginton (1980) was the first to apply logistic regression for credit scoring and reported that logistic regression was superior to LDA. According to Wiginton (1980) logit models are more suitable for credit scoring compared to other techniques such as DA because they do not assume that the variables are drawn from a multivariate normal distribution. In his study, he compared the performance of a binary logit model to a linear discriminant model using the same demographic variables. His results revealed that the logit model predictability dominated the linear discriminant results.

Advances in computing, have allowed other techniques to be adopted for building scorecards. The competitive environment in financial institutions and the growth of technology has made it possible for institutions to explore the use of non-parametric and intelligent techniques such as ANN, GA and k-nearest neighbor. Decision tree algorithms have also been broadly used in credit scoring resulting in good performance. GA was successfully used in credit scoring and the usefulness of this technique is discussed in a study by Goonatilake and Treleaven (1995). ANNs, due to their ability to identify non-linear relationships were used in credit scoring and many studies (Chuang & Lin, 2009; Desai, et al., 1996) suggest that they outperform other statistical credit scoring techniques for this reason. Presently, ANN is the most popular method used for credit scoring; however, its long training process in designing the optimal network topology and difficulty in identifying the relative importance of the input variables limits its application.

Table 2.7 considers the advantages and disadvantages of the most commonly used methods in credit scoring.
<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
</table>
| Linear discriminant analysis | 1. Easy to calculate.  
2. Available in many statistical software packages.  
3. Prior probabilities and misclassification costs can be easily incorporated into LDA. | 1. Linear discriminant analysis requires normally distributed data.  
2. Probability of default is not directly determined.  
3. Assumption of approximately equal variances in each group is not always met.  
5. Sensitivity to outliers. |
| Ordinary Least Square Regression | 1. Bigger flexibility, easy to use.  
2. The theory associated with linear regression is well-understood easily interpretable.  
3. It uses data efficiently. Good results can be obtained with relatively small data sets.  
4. The estimates of the unknown parameters obtained from linear least squares regression are optimal – (BLUE) Best linearly unbiased estimators. | 1. Output can be more than 1 or lower than 0.  
2. Linear regression assumes – Linearity, Homoscedasticity, and Normally Distributed error term which implies a continuous and normally distributed target variable; independent error terms and uncorrelated predictors.  
3. The extrapolation properties are poor.  
4. It is very sensitive to outliers.  
5. Calculation problems arise, when we deal with many variables. |
<table>
<thead>
<tr>
<th>Method</th>
<th>Advantages</th>
<th>Disadvantages</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic Regression</td>
<td>1. Output is directly determined</td>
<td>1. Logistic regression is sensitive to high correlation among explanatory variables; one should ensure that there are no such variables left in the training set.</td>
</tr>
<tr>
<td></td>
<td>2. No distributional assumptions on independent variables.</td>
<td>2. Sensitive to missing values (all observations with missing values have to be deleted).</td>
</tr>
<tr>
<td></td>
<td>3. It allows the assumption that dependent variable has a Bernoulli</td>
<td>3. It invites an over-interpretation of some parameters.</td>
</tr>
<tr>
<td></td>
<td>distribution, the appropriate distribution for binary data.</td>
<td>4. Calculation problems arise, when there are a lot of variables.</td>
</tr>
<tr>
<td></td>
<td>4. Scores are interpretable in terms of log odds. Probability of default</td>
<td></td>
</tr>
<tr>
<td></td>
<td>is always within than range [0, 1].</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. The logit method can deal with categorized data.</td>
<td></td>
</tr>
<tr>
<td>Decision Trees</td>
<td>1. It makes no distributional assumptions for dependent and independent</td>
<td>1. The relative importance of variables is not determined.</td>
</tr>
<tr>
<td></td>
<td>variables.</td>
<td>2. It is a discrete scoring system. Probability of default must be estimated using % bad in each terminal node.</td>
</tr>
<tr>
<td></td>
<td>2. It handles both categorical and continuous explanatory variables.</td>
<td>3. Sensitive to changes in the training data. Therefore decision tree is used in banking practice as a supporting tool.</td>
</tr>
<tr>
<td></td>
<td>3. It is a simple procedure method giving clear output, which is easily</td>
<td></td>
</tr>
<tr>
<td></td>
<td>interpretable.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>4. It can handle noisy and incomplete data.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>5. Decision tree method is not affected by outliers, collinearity,</td>
<td></td>
</tr>
<tr>
<td></td>
<td>heteroscedasticity, or distributional error structures that affect</td>
<td></td>
</tr>
<tr>
<td></td>
<td>parametric procedures.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>6. Decision trees have the ability to detect and reveal variable interactions</td>
<td></td>
</tr>
<tr>
<td></td>
<td>in the dataset.</td>
<td></td>
</tr>
<tr>
<td>Method</td>
<td>Advantages</td>
<td>Disadvantages</td>
</tr>
<tr>
<td>-----------------</td>
<td>-----------------------------------------------------------------------------</td>
<td>-------------------------------------------------------------------------------</td>
</tr>
<tr>
<td>Neural Networks</td>
<td>1. Model has no restrictive assumptions.</td>
<td>1. Not easy to interpret.</td>
</tr>
<tr>
<td></td>
<td>2. It is robust and flexible.</td>
<td>2. Complex calculation.</td>
</tr>
<tr>
<td></td>
<td>3. Interactions between variables are automatically assessed.</td>
<td>3. Requires high quality data, variables must be carefully selected a priori.</td>
</tr>
<tr>
<td></td>
<td>4. Can handle incomplete, missing or noisy data.</td>
<td>4. Requires a definition of architecture.</td>
</tr>
<tr>
<td></td>
<td>5. Can overcome autocorrelation.</td>
<td>5. Long processing time.</td>
</tr>
<tr>
<td></td>
<td>6. Capable of mapping any complex non-linear and/or approximate continuous function.</td>
<td>6. Large training sample is required.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>7. Expensive implementation and maintenance.</td>
</tr>
<tr>
<td></td>
<td></td>
<td>8. Big possibility of overfitting a model.</td>
</tr>
</tbody>
</table>
In today’s competitive world, the search for a good scoring model is still an ongoing process. Mixtures of techniques / combination of models and are often used to improve the accuracy of scoring model (Peng, et al., 2005; Zhu, Beling, & Overstreet, 2001) and this is the approach used in this study.

The purpose of combining models in credit scoring is to produce a final model that is better than the individual models in terms of accuracy rate (or other criteria) (Lee & Jung, 1999/2000). Zhu, et al., (2001) combined two sets of consumer credit scores in their study and proved that the combined score outperformed each single set of scores.

Combining models often results in improvement of model’s accuracy compared to the individual models performances. However, it is not easy to build a combined model which can significantly outperform the individual models and it gets more complicated to interpret the rules generated by the combined model. The extra time taken to develop and utilize combined models may not be reasonable if the improvement is minimal. In situations where one needs a quick measure of risk, combined models may not be very useful, but if there is no restriction on time and one is more interested in a better model, then combined models can be very valuable. It is always important to understand the purpose and the most important aims before building a scorecard, so that the right expectations can be set.

Building of scoring models has always been based on a pragmatic approach, therefore there is no gold standard and the best model differs for every unique situation. Each technique has its own limitation and advantages. The right technique depends on the data and the unique aims of the model.

Classification accuracy is not the only aspect of performance to be considered when building score cards. Other aspects such as speed of classification and interpretability should also be considered. For example, in terms of application scoring, an instant decision is much more appealing to a potential borrower than a decision for which the borrower needs to wait for several days. The same applies to behaviour scoring. Since time plays an important role in credit risk, in many cases techniques which are accurate enough in analysing the customer’s behaviour are preferred to techniques which perhaps
give a superior accuracy but take much longer to come up with decisions regarding the creditworthiness of customers. Some of the techniques, despite having good classification accuracy, are not commonly used because of computational problems or interpretation complexity. Classification methods which are easy to understand and interpret such as regression and decision trees are more appealing compared to methods such as NNs which are regarded as black-box techniques without rule and logic, which are not easily interpreted.

Davis et al., (1992) compared the accuracy of several techniques and concluded that overall, all the techniques perform to the same accuracy with only the NN algorithm having a longer training time.

Hand and Henley (1997) review various classification methods in credit scoring and concluded that there is no overall ‘best’ method. They also refer to their comparison work (1996) where they developed an adaptive metric nearest neighbour method with a parameter describing the shape of the metric (D) for credit scoring. In this study nearest neighbour methods were shown to be superior compared to linear regression, logistic regression and decision tree. However Hand & Henley (1997) highlight that the differences between the performances of the methods are not of any practical value if the cost of changing a scoring system and the likely future life of any system that is installed are considered. According to these authors what is best depends on the particulars of the problem, that is, on the data structure, the characteristics used, and the separation of the classes by using those characteristics and the objective of the classification. They believe that the significant improvements in accuracy of classification are more likely to come from including new predictive characteristics than from new techniques. This is the approach followed in this thesis.

In this research we have aimed to include new predictive characteristics in our model and we apply various techniques and compare the performance of these techniques in this context.


2.9  Limitations of credit scoring:

Model based credit scoring has many advantages over judgmental methods. Credit scoring systems not only make granting credit a much quicker process, but they also make the process more consistent and they improve the reliability of decisions about whether to grant credit to an applicant. Most importantly, scoring systems reduce social discrimination by excluding non-risk related variables such as sex, race and religion. By automating the process, the financial institutions can deal with a much larger number of applicants. Credit scoring systems are also beneficial for customers since they are able to know the outcome of their application in a few hours rather than days. For example, for a business loan, the approval processing time has been reduced from two weeks to 12.5 hours (Allen, 1995). However, this very positive advantage of credit scoring systems may reduce the amount of interaction between the banks and their customers, which may possibly contribute to a negative image for the banks (Matsatsinis, 2002).

All credit scoring models handle quantitative attributes well; however, qualitative attributes are not easily handled by some scoring models (Henley, 1995).

A major drawback of credit scoring systems is the change of pattern over time. This is called population drift. The impact of population drift is of interest to both academics and practitioners (Hand & Henley, 1997). Population drift occurs because populations change over time. This change could cause the changes in distribution of population characteristics and hence model’s prediction capability could diminish. The key assumption of credit scoring models is that the future is like the past (Berry & Linoff, 2000). This assumption is not always valid due to changing economic conditions and changing customer perceptions on borrowing (Thomas, et al., 2004). Hence credit scoring systems need to be adjusted frequently to consider new situations.

In regards to behaviour scoring, one of the disadvantages is that it typically needs two years transactional history to build scorecards. The population on which the scorecard is built is quite different to the population to which it is applied (Thomas, Ho, & Scherer,
Another common problem in credit scoring is bias in the sample data from which the credit scores are built (Hand, 2001). Mostly the information available is of the approved applicants. This data is not appropriate for developing models for identifying “good” or “bad” customers from their posterior behaviour. There is no information on the rejected applicants which makes the available sample data biased.

Another problem of credit scoring models is the proportion of good and bad risk customers; what proportion of good credit customers and bad credit customers should be allowed in the sample used to train the model? This problem is addressed in this thesis by oversampling and clustering. Berry and Linoff (2000) explain that with a binary outcome the ratio of 20-30 percent of rare cases often produce good results. We over-sample the rare cases (future bad customers) to get 20% future bad customers in the training data. Also we cluster the customers into smaller groups of similarly behaved customers. In each group we have a good proportion of future good and future bad customers.

Yet another problem is the omission of important variables in the model (Avery, Bostic, Calem, & Canner, 2000). Credit scoring models are based on individuals’ credit and payment history. This may not be enough to come up with a conclusion of credit worthiness for an individual. An individual is classified as a “bad” customer if his characteristics are similar to the characteristics of a bad customer. However, default could be a cause of economic status or unemployment which means that these characteristics cannot be used to predict future default. Nonetheless, by using only one month’s behavioural data to build our models for delinquency it is expected that this problem will be overcome.
Chapter 3
DEVELOPING A CREDIT SCORING SYSTEM

But in the new approach, as you know, the important thing is to understand what you’re doing rather than to get the right answer

- American mathematician, singer, songwriter Tom Lehrer
3.1 Introduction

This chapter introduces the methodology and description of the methods selected for this thesis. The new Basel II Accord allowed banks to base their capital requirements on their own internal credit risk rating system. After the pronouncement of Basel II, banks and financial institutions gave great importance to credit risk models. The financial institutions are working harder to develop new innovative credit risk models and to improve the accuracy and usefulness of existing models.

With the new Basel regulations, banks, for the first time, are permitted to rely on their own assessments of a borrower’s credit risk. This means that changes in a customers’ credit behaviour will directly influence the amount of capital held by banks (Angelini, et al., 2008). This gives banks more control of their risk and consequently gives more
weight to the management disciplines of measuring, monitoring and controlling risk (McDonough, 2003). For this reason the search for more sophisticated and more accurate credit scoring systems is ongoing.

This thesis aims to develop a new approach to credit scoring. Many financial institutions are interested in applying different approaches and comparing the new approaches to existing approaches, but time constraints make it impossible for this to be done in-house. In addition the thesis develops an approach for tracking changes, producing information which can be used by banks to provide a better service to their customers.

We have used Enterprise Miner version 4.3 in SAS to fit all our models while SAS version 9.1 has been used for all data manipulation. The main reason for using SAS for all the analysis is because SAS can handle very large data sets efficiently.

As mentioned in the previous chapter, there are 6 steps in developing a scorecard; 1) Data preparation, 2) Data cleaning, 3) Variable selection, 4) Sample generation, 5) Model development and validation, 6) Model approval. Each of the steps is explained in detail in the following section.

In section 3.2, we discuss the data processing steps which involve the data preparation and cleaning, variable selection techniques and problems involved in generating a good sample for developing a credit scoring system. In section 3.3 various methods that are used in this research for modeling are discussed in detail, with their limitations and advantages. Section 3.4 describes the performance measures which measure the accuracy of the classifier. Table 3.1 provides the outline of this chapter.

### 3.2 Data processing:

Credit scoring is based on historical data. Quality of the data is an important factor in building a successful scorecard. If irrelevant/redundant information or noisy/unreliable data is used, building a credit scoring model becomes more difficult or one may get
unreliable/biased results. Therefore a good understanding and careful preparation of the dataset is essential for developing good credit scoring models.

3.2.1 Data Preparation and Data Cleaning:
As said by Capon (1982), it is a mistake to use any variable that can improve the performance statistically, each variable must have obvious relevance for predicting risk. A good understanding of these variables and careful preparation of the data is a necessary step in developing a successful credit scoring model.

Raw banking datasets more often than not contain hundreds of variables. One needs to select the most relevant and significant variables according to the purpose of the scorecard. Sometimes new variables are created from the raw data to reveal hidden information for credit scoring. According to Hand and Henley (1997) the significant improvements in accuracy of classification are more likely to come from including new predictive variables than from new techniques.

One of the major contributions of this research is in regard to data preparation. This research aims to create new variables, which we believe indicate the financial circumstances of the customers and are good predictors of behaviour.

Data cleaning helps detect and correct any corrupt and inaccurate information which may bias the dataset. In the data cleaning process, the variables which may not legally be used in credit scoring are removed from the dataset. In Australia one such variable is the gender of the customer (Leung, Cheong, & Cheong, 2007). In the United Kingdom, it is also illegal to use variables such as race or sex in credit scoring (Thomas, et al., 2002). Also it is good practice to remove variables with a high percentage of missing values and variables with little or no variability.

3.2.2 Variable Selection
After discussion with analysts from financial institutions it was concluded that data cleaning and variable selection are the most time consuming activities in credit scoring. Banking data are large and complex. As stated by Bellman (1961), along with complex data sets comes “the curse of dimensionality”. Higher dimensionality means more
variability and more parameters needed to specify the distribution. To be able to analyse and understand complex data sets, dimension reduction methods are often necessary to select the significant features of the data, to discard the background noise and to make interpretation possible (Breiman, Friedman, Olshen, & Stone, 1984).

Variable selection is of extreme importance in scorecard development. Banking data are multi dimensional with a large set of predictor variables. Not all the variables are useful in the field of credit scoring. Typically after cleaning the data, the number of variables available for credit scoring ranges from 50-60. Using such a large range of variables would not only result in over-fitting (Sanche & Lonergan, 2006) but would also reduce the efficiency of the score card. Selecting only a small subset of variables results in scorecards that are easier to understand, and removing irrelevant information decreases the computational time and storage requirement, which saves time and money (Guyon & Elisseeff, 2003; Serendero & Toro, 2003; Zhang, 2000). Therefore it is best to have parsimonious scorecards with only a small number of relevant variables, resulting in better credit decisions. Fair, Isaac & company (Fair, Isaac & Company, Inc. 1996) found that, out of 50-60 available variables, only 8-12 variables were needed for their scorecard. Hence there is a need to identify the most effective combination of predictor variables and to then remove irrelevant and redundant information to enable better prediction. This selection of variables can be conducted automatically with binary logistic regression methods in SAS.

There are three different logistic regression methods for variable selection. These are iterative approaches and involve considering only addition or deletion of a single variable at each step.

1. Backward elimination:

This method starts with all available explanatory variables in the model. It fits the best possible model by dropping one non-significant variable at a time. The variable to be deleted is the variable whose coefficient has a $t$-statistic closest to zero. This process continues until all remaining $t$-ratios are above some predetermined level usually corresponding to the 10% or 5% significance levels. Once a variable is dropped from the model, it cannot enter the model again. Backward elimination is generally to be preferred if it is computationally feasible,
as this allows the inter relationships between all the variables to be assessed before any variable is removed.

2. Forward elimination:
   This procedure starts with an initial model which contains only the intercept and no explanatory variables. One variable is added to the model at a time. The first model contains the intercept and a single explanatory variable which creates the model. Then the remaining variables are added to the initial model to see which one creates the model. This continues until the t-ratio for adding another variable is not significant at a predetermined level. Once a variable has entered the model it cannot be removed at a later stage. The main advantage of forward selection over backward elimination is that the initial models that are fitted only contain a small number of explanatory variables. This means that this approach is less time consuming when there are a large number of possible predictor variables.

3. Stepwise regression:
   This is a combination of backward elimination and forward selection. The procedure is the same as forward selection except that after each step of forward selection, a backward elimination step is attempted. A higher significance level ($\alpha$) is normally used for adding variables than for deleting them, in order to keep useful variables in the model.

3.2.3 Sample Generation
Since credit scoring datasets are usually very large and multidimensional, it is common to take samples for scorecard development rather than the full population. This is helpful in saving time to pre-process the data, model and report. Also it saves computer storage capacity. Sample generations faces two problems; what sample size to select and what proportion of future good and bad customers to include in the sample?

i. Sample size
There is no rule for choosing the right sample size. It is found in the literature that it is common to use 1,500 – 2,000 cases of each class (good/bad) to develop good quality models (Lewis, 1992; Siddiqi, 2006). As said by McNab and Taylor (2008) there is little statistical benefit in taking samples of more than 5,000 cases for each class. Unnecessarily large sample sizes increases running time, making the development of an optimum method problematic. However, one benefit of using a large enough sample is that it reduces the impact of multicollinearity (Achen, 1982).

ii. Data imbalance
A credit scoring model performs well when the model is applied to a dataset evenly distributed among different classes. However, many real world credit scoring datasets involve imbalanced class distributions i.e., there are many more customers in one class compared to the other class. In such cases, a classifier usually tends to predict that customers have the majority class label, completely ignoring the minority class, therefore biasing the predictions (probabilities) in favor of the dominant class. In cases such as scorecards, recognizing the minority class (bad credit customers) is very important since it helps the organizations take appropriate actions to minimize their risk/loss. Therefore one needs to select a sample which overcomes the imbalance in the data. This can be addressed by re-weighting the data (Domingos, 1998) or by over-sampling (Chawla et al., 2002, 2003; Japkowicz, 2001) or under-sampling (Chyi, 2003; Zhang & Mani, 2003). Over-sampling, in its simplest form, duplicates rare cases (minority-class) while under-sampling reduces majority-class cases (Burez & Van den Poel, 2009). These sampling methods help in decreasing the overall level of class imbalance; however, they have several drawbacks (Weiss, 2004). Under-sampling removes useful majority-class examples and thus can downgrade the performance of the classifier. Over-sampling duplicates the minority class and often involves making exact copies of examples, which may lead to overfitting (Chawla et al., 2002; Drummond and Holte 2003). Also because over-sampling introduces additional cases it results in increasing the time required to build a classifier (Burez & Van den Poel 2009). Also over-sampling does not introduce new data, it only duplicates the minority cases, and therefore it does not tackle the fundamental “lack of data” issue. This clarifies why simple over-sampling has been shown to be ineffective at improving recognition of the minority class (Ling and Li, 1998; Drummond and Holte 2003) and why under-
sampling seems to have an advantage over over-sampling (Chen et al. 2004). For these reasons in this research we use under-sampling. In the main study there were less than 8% ‘bad’ customers. Under-sampling was performed including all the “bad” customers and selecting good customers randomly so that we had a sample containing 20% bad customers and 80% good customers. Under-sampling is performed so that the rare class becomes less rare in the sample, therefore reducing the effect of imbalance and increasing the accuracy of the model predictions. Classifiers which ignore the imbalanced class distribution problem lose their ability to predict the minority class correctly (Yen & Lee, 2009).

The dataset used in this research represents a credit card portfolio which is imbalanced, that is, the number of bad customers is small (minority class) compared to the number of good customers (majority class). In this research additional to under-sampling, clustering is performed before variable selection. In this research customers are clustered using a Self Organizing Map (SOM), an unsupervised network. The SOM groups the customers into clusters based on similarity in terms of the significant input variables (the SOM selects the important variables which have a significant role in clustering). Three clusters with the highest percentage of future bad customers are selected for modeling one at a time. Clearly the imbalance problem is reduced for theses clusters because they are selected on account of their high percentage of future bad customers. Different techniques were used for each cluster and the best (minimum misclassified rate) was selected as the preferred model for that particular cluster. By doing this, we pick different significant characteristics for each cluster. The preferred model developed for each of these clusters are then applied to the whole dataset, effectively producing additional predictor variables while using a dimension reduction approach.

Yen and Lee (2009) have proposed a few under-sampling methods based on clustering and have shown that using under-sampling, the influence of an imbalanced class distribution can be decreased and the accuracy of predictions for the minority class can be increased.

iii. Data Partition

Partitioning data is an important task for validating data mining models. When dealing with known inputs and corresponding known outputs, we evaluate the model by
splitting the dataset into at least two parts (e.g., training and test data). The training dataset is used for fitting the model and the test dataset is used to evaluate the accuracy of prediction. Dividing the population into training and test datasets should be done randomly so that the test error rate is a reliable estimation of the true error rate. Typically, more than half of the dataset is used for training. This leaves a smaller dataset for testing. If the dataset is small originally we cannot expect accurate results from this approach, also it is possible to get an “unfortunate” split, using random sampling resulting in misleading results. In these situations a more computationally intense validation technique such as cross-validation should be used to estimate the prediction accuracy of the model. If the dataset is large enough it is better to divide the data into training, validation and test data (see Figure 3.1).

The data provided to us for the main study was very large and therefore we partitioned the data into three parts. The training data consists of 40% of the data which is used for learning and for preliminary model fitting. The validation data set (30%) is used to monitor and tune the model weights during estimation and is also used for model assessment. The remaining 30% of the data is an additional hold-out data set known as test data which provides a measure of predictive accuracy. The goodness of prediction (error rate estimate) of the final model on the validation dataset will be biased (i.e., smaller than the true error rate) since the validation dataset is used to select the final
model. Therefore having a separate validation and test dataset is useful. An important fact to keep in mind is that the test dataset should never be used for training or tuning the model.

3.3 Model Development and Validation

Once the samples are generated and the data has been partitioned into training, validation and test (holdout) samples, models can be developed. The training set is used for initial modeling and the validation set is used to tune the parameters of the model. The type of tuning employed depends on the type of classification technique that is used. There are several classification techniques for model development including traditional statistical as well as intelligent systems techniques.

Linear discriminant analysis (DA) was the first technique used for building credit scoring model (Reichert, et al., 1983) and is the most popular method for building credit scoring models in the literature. According to many analysts, Logistic Regression (LR) is the most common technique used for building credit scoring model. Decision tree algorithms have also been broadly used in credit scoring resulting in good performance (Boyle, et al., 1992; Srinivasan & Kim, 1987). Competitive environment, rapid growth of technology and the dramatic growth of credit users brought in front the need for a safer and more accurate evaluation of credit scoring system. This led to the implementation of sophisticated credit scoring systems such as non-parametric and intelligent techniques such as Artificial Neural Network, Genetic Algorithm and k-nearest neighbor. In today’s competitive world, the practitioners and researchers are still exploring and testing new and more complex scoring techniques.

In addition, mixtures of techniques and hybrid models are used to improve the accuracy of scoring model (Hsieh, 2004; Peng, et al., 2005; Zhu, et al., 2001) and this is the approach used in this study. All the methods used in this thesis are explained in detail in the following chapter (chapter 4). In this research we have used a hybrid credit scoring model based on clustering and mixture of algorithms (decision tree, logistic regression...
and ANN). A two-stage model is used to predict probability of a future bad customer at first stage and the amount of balance owed at the second stage using logistic regression and regression respectively. When the model is developed, the probability of customers becoming delinquent is calculated and based on these probabilities customers are recognized as ‘good’ or ‘bad’ class using an appropriate cutoff. However, it is also a common practice to work with this probability after converting it to a score ranging from 100 to 1,000.

Thomas et al., (2002) believe that converting probabilities into scores of 100 – 1000 has psychological and operational effects. They believe that it is easier for people to understand big round numbers that range from 100 to 1,000 rather than probabilities between 0 and 1.

Most financial institutions have their own standard alignment scale which they use to convert the probabilities to scores. However, due to confidentiality issues, the banks which provided the data to us could not provide us with this scale. Therefore this study does not use any scores; instead it refers to the probabilities directly.

Once the probability of delinquency of all the customers has been generated for the training dataset, the model accuracy should be checked. Model accuracy is calculated by performance measures such as Percentage Correctly Classified (PCC), Gini or ROC curve. The credit scoring model is validated by applying the same process to the test (holdout) sample (that is, applying the selected model to test data, generating default probabilities and then calculating the performance of the model for the test sample). The various performance measures are explained in the following section.

3.4 Performance Measures

In behaviour credit scoring, the aim is to produce a score which represents the probability that an existing customer should be classified as either a “good” or “bad” future customer. As said in the introduction chapter, a “good” future customer is a
customer who has no overdue payments and a “bad” future customer is a customer who has is minimum 90 days overdue.

The performance of a classification model is measured based on the classification accuracy of the model in the case of the test data. In other words, the performance is determined by how well the model can correctly classify the bad customers as bad and good customers as good. In many cases, the cost of misclassification is high (or very high), therefore the accuracy of classification models needs to be thoroughly evaluated. Most credit risk practitioners in financial institutions prefer the Gini coefficient as a measure of performance in the field of credit scoring. In the field of credit scoring, the Gini coefficient is calculated based on a credit score. Although the Gini coefficient is preferred in financial institutions, only a few studies use the Gini coefficient in the current literature. The ‘percent correctly classified’ is more commonly used as the primary measure for assessing the predictive performance of scorecards. Some of the other features that could be used are the gmean, F-measure, Kappa statistic, and receiver operating characteristics (ROC) curves along with their associated area under the curve. All these performance measures are explained below.

In this thesis we have used percent correctly classified, ROC curves, and the Gini coefficient. However the preferred measure of performance is the ROC curve and the Gini coefficient.

Credit scoring is a binary decision problem. A selected classifier labels each customer as either ‘good’ or ‘bad’. This is represented in a structure below known as the confusion matrix, which shows the predicted and actual delinquency level for all customers. A confusion matrix is of size m ×m, where m is the number of different classes. In credit scoring the number of classes is two (“good”, “bad”), therefore we have a 2x2 confusion matrix which is presented in Table3.2.
The confusion matrix has four categories:

- True Positive (TP) are ‘good’ instances correctly identified as ‘good’;
- False Positive (FP) refers to ‘bad’ instances incorrectly identified as ‘good’;
- True Negative (TN) corresponds to ‘bad’ instances correctly identified as ‘bad’; and
- False Negative (FN) refers to ‘good’ instances incorrectly identified as ‘bad’.

- \( N \): total number of instances where \( N = TP + FP + TN + FN \);
- \( G_T \): total instances in ‘good’ class;
- \( B_T \): total instances in ‘bad’ class;
- \( G_p \): predicted instances in ‘good’ class;
- \( B_p \): predicted instances in ‘bad’ class.

### 3.5.1 Percent Correctly Classified (PCC)

This commonly used measure is the proportion of correctly classified instances from the confusion matrix which includes correctly classified good customers and correctly classified bad customers. Thus, this measure is calculated as the sum of true positives (TP) and true negatives (TN) over the total number of instances (N).

\[
PCC = \frac{TP + TN}{N}
\]

The assumption behind this measure is that correctly classifying bad customers and correctly classifying good customers have equal importance. But this is not true in the
credit scoring problems, since correct prediction of bad risks is more important due to the associated higher cost of misclassifying a bad risk. This measure is particularly misleading in a highly unbalanced credit scoring dataset where the percentage of ‘bad’ instances is very low (say 6%) compared to the percentage of ‘good’ instances (say 94%). If the classifier classifies all the instances as good, we get only 6% misclassified and PCC is 94% which is very high. This leads us to a conclusion of a highly accurate classifier, while this measure fails to classify any bad cases correctly. This measure of performance therefore has little value in practice (Freeman & Moisen, 2008). However, in the current classification literature, many studies still use this measure for comparing the performance of classifiers, perhaps because of its simplicity.

3.4.2 Receiver Operating Characteristics (ROC) Curves

The most important characteristic of a classifier lies in its ability to distinguish between the states (good/bad) taken by the dependent variable. But the actual classification results depend on the threshold to which the predicted probabilities for each outcome category are compared. Thus there arises a need for a measure of discrimination that is not dependent on arbitrary choices of threshold. One solution to this problem is provided by the Receiver Operating Characteristics (ROC) curve.

Receiver operating characteristics (ROC) curves are a useful technique for assessing classifiers and visualizing their performance. ROC curves are insensitive to changes in class distribution and are therefore unaffected by the presence of unbalanced classes, which is very common in real-world credit scoring datasets. This makes ROC curves appropriate for credit scoring datasets. In addition to being a useful graphical method, ROC curves are used in determining if a classification model is working beyond pure chance and, also, in recognizing which classification model is better at distinguishing between groups (Hall & Mayo, 2005). The ROC curve does not assume a single threshold and therefore does not actually do any classification. ROC curves can only be obtained for classifiers that can produce a probability or score. Discrete classifiers (those which are able to produce only a class decision) are not capable of producing ROC curves. This is because these classifiers generate only one ROC point, whereas
classifiers which produce a probability or score generate a ROC point for each possible threshold.

Before going further it is important to define the two basic concepts of sensitivity and specificity for a classifier; sensitivity and specificity. *Sensitivity* measures how often we find what we are looking for (say, bad customers), whereas, *specificity* measures how often what we find is what we are not looking for. In other words, Sensitivity ($S_e$) is the proportion of actual negatives (bad customers) which are correctly identified as negatives by the classifier, that is

$$S_e = \frac{TN}{BT}$$

The Sensitivity is 1 if all negatives (bad customers) are correctly identified.

Specificity (Sp) is the proportion of actual positives in the data (good customers) which are correctly identified as positive by the classifier.

$$S_p = \frac{TP}{GT}$$

Specificity is 1 only when all positives are correctly identified.

A ROC curve plots sensitivity against (1 – specificity). In other words, this curve plots the percentages of correctly classified negatives (Y-axis) against the percentages of incorrectly classified positives (X-axis) for each possible prediction threshold. Each point on the curve depicts classification for a particular threshold where the distance below the point represents the proportion of true negatives correctly classified and distance to the right represents the proportion of true positives incorrectly classified.

The area under the curve can be viewed as the proportion of correct classification across all the possible thresholds. This area ranges between 0 and 1 and gives a good summary statistic of discrimination. Figure 3.2 depicts an example for several ROC curves. The straight line passing through points (0, 0) and (1, 1), is the reference line which represents a classifier whose performance is not better than random guessing and which
therefore has poor discriminatory power. Model D in Figure 3.2 is a perfect model. The more an ROC curve approaches point (0, 1), the better is the classifier (for example, Model A outperforms Model C in Figure 3.2).

![ROC Curves](image)

**Figure 3.2** ROC curves

ROC curves show a graphical evaluation of classifiers. However, when the curves intersect each other, this evaluation becomes less obvious (see Models A and B). To have a clear view, it would be better to have a single value representing the expected performance. A common method of doing this is calculating the *Gini coefficient* (Breiman, et al., 1984) which measures the models’ ability to separate risk. As pointed out before, the Gini coefficient is the preferred measure of performance in the field of credit scoring since it is able to rank risk throughout the entire set of credit scoring data, without giving any special weight to performance near the accept/reject region (Burns & Ody, 2004).

The Gini coefficient is calculated as (Hand & Till, 2001):

\[ G = (2AUC - 1) \]

where AUC is the area under an ROC curve (Bradley 1997, Hanley & McNeil 1982). The AUC has a significant statistical property in that the AUC of a classifier is equivalent to the probability that the classifier will rank a randomly selected negative
observation (bad customers) higher than a randomly selected positive observation (good customer) (Fawcett, 2003). The maximum value of the axes for the ROC curves is always one, therefore the value of AUC will always be between 0 and 1 (refer Figure 3.2) Random guessing, shown as the reference line in Figure 3.2, produces an AUC of 0.5 and a Gini coefficient of zero.

For a behaviour scoring models Gini value of 0.65-0.8 indicated a good credit scoring model. Table 3.3 shows the quality of classification by the Gini coefficient.

<table>
<thead>
<tr>
<th>Gini value</th>
<th>Classification Quality</th>
</tr>
</thead>
<tbody>
<tr>
<td>&lt;0.45</td>
<td>Low</td>
</tr>
<tr>
<td>0.45-0.64</td>
<td>Average</td>
</tr>
<tr>
<td>0.65-0.80</td>
<td>Good</td>
</tr>
<tr>
<td>&gt;0.80</td>
<td>Very Good</td>
</tr>
</tbody>
</table>


3.5 Model Approval

In practice, the last step in building a scorecard is model approval. The financial institution must approve the selected credit scoring model according to the business rules of the financial institutions. Sometimes small changes are made to the scorecard to meet the requirements of the business rules. One such example is changing the coarse classing standard to reflect business practice. Although the data for this research was provided by a very large Australian bank and the institution was very interested in this research, this step has yet to be taken.
the3.6 Summary

This chapter has described the steps involved in building a credit scoring model. These steps are 1) Data preparation, 2) Data cleaning, 3) Variable selection, 4) Sample generation, 5) Model development and validation, 6) Model approval. In the next chapter we describe the specific methodology that will be used in this research.

The aim of this research is to apply methods commonly used in industry today, explaining all the steps involved in building credit scoring models and considering practical ways of improving the performance of these models.

As said before the cost of changing scoring systems is high (Hand & Henley, 1996) and banks prefer to use well known methods of credit scoring that are easy to understand. The main focus of this research was to develop tools which are readily available to banks. Therefore newly developed methods such as Support Vector Machine (SVM) and Random Forest (RF) are not explored in this research since they require additional implementation costs and are not so easy to understand.

In summary the objectives of this research are highlighted below;

- Create appropriate measures
- Create clusters of customers with similar spending behaviour
- Track behaviour changes and predict risk for individual customers
- Develop models for risk using short time series
- Validate models for risk using out of sample data
Chapter 4

METHODOLOGY

I AM ON THE VERGE OF MYSTERIES AND THE VEIL IS GETTING THINNER AND THINNER.

-FRENCH MICROBIOLOGIST LOUIS PASTEUR (1822-1895)
4.1 Introduction

In the previous chapter three important steps involved in credit scoring system were introduced. These steps were data processing, model development and model validation. This chapter explains the techniques used to implement these steps in this research. The outline of the chapter is given in Table 4.1.

4.2 Outline of the studies

This thesis includes two studies. First a pilot study was performed on a smaller set of banking data. In this analysis 6 months of transaction data were considered for 6000 customers. These customers were classified into different groups using hierarchical clustering based on transaction amounts in order to help the bank to distinguish between big spenders and small spenders. This is an important task for banks in terms of marketing strategies and risk. By differentiating between big spenders and small spenders, banks can identify a group of customers who are more profitable; allowing the
banks to offer these customers other services and have reward programs, helping the bank to retain this profitable group of customers. On the other hand, they also need to be aware of small spenders and to come up with strategies to encourage them to spend more. Also in this study the customers were grouped according to the products they buy, allowing the bank to target each specific group of purchasers with appropriate banking products. The behaviour of the customers was analyzed each month and the movement from one cluster to another was monitored, allowing the banks to better understand when/if there were changes in the financial circumstances of their customers. Monitoring the behaviour of customers over time allows customers’ financial movements to be detected. This movement could be used to identify customers moving from good clusters to bad clusters, or vice versa and for identifying steady customers who are in same clusters for a long time. When such movements are observed over time banks can come up with appropriate strategies to deal with individual customers. Also this analysis shows what customers are buying. This is useful to the bank in terms of marketing new products.

For the second analysis the transactional data, delinquency status and balances of approximately 1.5 million customers were considered. The purpose of this analysis is to create variables which we believe are significant in identifying risky customers and also to combine credit scoring models to get better accuracy compared to the individual models. For this analysis the transactions were separated from the payments with fees and fines excluded. Each customer’s transactional data was aggregated to a daily level, and divided into a separate monthly time-series for each account. steps involve in this study is as follows;

a) The analysis started with the fitting of an AR (7) model to each customer’s series. Autoregressive parameters (Ø1 – Ø7) were estimate with lag 7 associated with weekly patterns of expenditure.

b) Moments (mean, standard deviation, kurtosis, and skewness) were then calculated for each time series.

c) Product based variables were created. The percentage spend on each product was calculated and treated as an input variable. These product based variables along with AR parameters and moments form the inputs for the clustering of accounts.
Based on these variables, accounts were clustered using self organizing maps, an unsupervised neural network. Clusters with a high percentage of future delinquent customers were selected for modeling.

Logistic regression, Tree classification and neural networks were applied and compared for modeling delinquency in each of these clusters. The best classification rule was chosen for each cluster.

Using these predictions as well as all the original variables in the training data two stage modeling was performed to estimate the delinquency probability for each customer first, and then the amounts of money the customers would owe were predicted at the second stage, so that the bank could differentiate between high and low risk defaulting customers.

The results were compared for various forecast horizons. Figure 4.1 explains the steps involved in the two studies. These two studies are explained in the following chapters.
The following section describes the various methods used in this thesis. We start by explaining the methodology used for our exploratory analyses (clustering and Markov Chains).

4.3 Clustering:

Clustering is the process of organizing objects into groups whose members are similar in terms of the clustering variables. Clustering is probably the most important unsupervised learning procedure. It deals with finding a structure in a collection of unlabeled data. So, the goal of clustering is to determine the intrinsic grouping in a set of unlabeled data.
The appropriateness of any cluster solution depends on the aims. In this thesis we use cluster analysis to improve the discrimination between good and bad customers. Many authors have used cluster analysis in this way (Hsieh, 2004; Lundy, 1992; Yen & Lee, 2009). However, Banasik et al., (1996) have suggested that if the segments are not distinct enough, segmentation may not always improve the scorecard. We consider three clustering approaches – hierarchical, Kohonen’s SOMs and time series clustering.

4.3.1 Hierarchical clustering
Hierarchical clustering produces a nested hierarchy of similar groups of objects based on the pairwise distance matrix for the objects. Hierarchical clustering is widely used since it does not require the user to specify any parameter, such as the number of clusters, and it does not make any assumptions about the data distribution. Though this method has great visualization appeal (Keogh & Kasetty, 2002; Mantegna, 1999) its application is limited to small data sets (Keogh, Lin, & Truppel, 2003). In the first analysis, Ward’s method was applied to a relatively small dataset of 1000 customers for 6 months of daily transactional data.

Ward’s method is a minimum distance hierarchical method, which calculates the sum of squared Euclidean distances from each case in a pair of cluster to the mean of all variables for these clusters. The pair of clusters to be merged at each stage has the lowest sum of squared Euclidean distances. There are several methods available for choosing how many clusters are required (for example, Cubic clustering criterion, Pseudo F, Pseudo T) but the best approach is to choose the number of clusters so as to ensure that the clusters are sensible and useful. PROC CLUSTER was used for this purpose.

a) Cubic Clustering Criterion:

The Cubic Clustering Criterion (CCC) was developed by SAS (Sarle, 1983) as a measure of the deviation of the clusters from the distribution expected if data points were obtained from a uniform distribution. The Cubic Clustering Criterion (CCC) is
used to estimate the number of clusters using methods which, like Ward’s minimum variance method are based on minimizing the within cluster sum of squares. The Cubic Clustering Criterion is obtained by comparing the observed $R^2$ values to the approximate expected $R^2$ values, assuming that there are no clusters, only uniformly (evenly) distributed data points using an appropriate variance stabilizing transformation. In this formula $R^2$ is the between cluster sum of squares divided by the total sum of squares.

The criterion is calculated as

$$\text{CCC} = \ln \left[ \frac{1 - E(R^2)}{1 - R^2} \right] \times K$$

where $R^2$ is the observed percentage variation explained, $E(R^2)$ is the expected $R^2$ assuming a uniform distribution, and $K$ is the variance-stabilizing transformation (Sarle, 1983).

Clusters are expected if the obtained $R^2$ value is greater than what would be expected if sampling from a uniform distribution. In this situation the CCC values are positive. Larger positive values of the CCC show a better solution, as it indicates a larger difference from a uniform distribution. One should keep in mind that the CCC may be incorrect if clustering variables are highly correlated.

In SAS this value is obtained using PROC FASTCLUS.

Values of the cubic clustering criterion between 0 and 2 indicate potential clusters, values greater than 2 or 3 indicate good clusters, and large negative values can indicate outliers. The CCC is calculated for cluster solution involving 1, 2, 3, etc. clusters. A peak in the plot of the CCC against the number of clusters is used to suggest the optimum number of clusters.

b) Pseudo-\textit{F} Statistic:

The pseudo-$F$ statistic is the ratio of the mean sum of squares between clusters to the mean sum of squares within clusters and is used to measure the 'tightness' of clusters.
(Lattin et al. 2003). In SAS this value is obtained from PROC FASTCLUS and is calculated as

\[
Pseudo - F = \frac{(T - P_G)/(G - 1)}{P_G/(n - G)}
\]

where \( G \) is the number of clusters, \( T \) is the total sum of squares, \( n \) is the sample size, and \( P_G \) is the within-cluster sum of squares. Larger numbers of the pseudo-\( F \) indicate a better clustering solution. Again plots of this statistic against the number of clusters are used to suggest the optimum number of clusters. Relatively large values for the pseudo- \( F \) indicate the optimum cluster number.

4.3.2 Kohonen’s self organizing network / Self Organizing Maps (SOM):

The Self-Organizing Map (SOM) acts both as a projection method which maps high dimensional data into low dimensional space (usually one or two dimensions), as well as a clustering method which maps similar data samples to nearby neurons (clusters). SOM is a powerful neural network method for the analysis and visualization of high-dimensional data. It was developed by Prof. Teuvo Kohonen in the early 1980s and has been used in a variety of applications such as visualization of high-dimensional data (Vesanto & Alhoniemi, 2000), identification of fraudulent insurance claims (Brockett, Xia, & Derrig, 1998), failure prediction (Huysmans et al, 2006) and many others. A broad overview of successful applications is presented by Kohonen and Deboeck (Kohonen, 1995; Deboeck & Kohonen, 1998).

SOMs are a type of unsupervised learning which aims to discover the underlying structure of the data in such a way that similar objects are assigned to the same or nearby nodes (Kohonen, 1995). There is a topological structure imposed on the nodes in SOMs, so they are called topology preserving maps. A topological map is simply a mapping that preserves neighborhood relations. SOMs are widely used as a data mining and visualization method for complex data sets, due to their robustness in regard to variable selection, their natural clustering results and their superior visualization for
very large datasets. Another advantage of SOMs is that they are less likely to get stuck in bad configurations (akin to local optima) for highly nonlinear data. However, very little research has been done to explore their usefulness for credit scoring (Huysmans et al 2006)

In SOMs, the nodes (neurons/clusters) are organized into a grid (usually a two dimensional grid). The grid space is separated from the input space. Any number of input variables can be used as long as the number of inputs is greater than the dimension of the grid space. Input variables can be binary, nominal, ordinal and interval. However, having many categorical variables, results in a longer execution time since they have to be converted into binary variables for all but one of the categories. The Kohonen self-organizing map algorithm enforces lateral interconnection directly by restricting weight adoption to topological neighborhoods. Each node is associated with a weight vector of the same dimension as the input data vector. The initial weights are randomly assigned to each node. The training starts with competitive learning. During the training process, a series of input vectors (data points) are fed into the network several times. When a new input vector is fed to the network, each node compares its weight vector with the input vector. A Euclidean distance to all weight vectors is computed and the node with most similar weight vector to the input vector wins. This node is called as winning node or Best Matching Unit (BMU). The weight vector of the winning node and the nodes close to it (neighborhood set) are then adjusted to better match this input vector and hence becomes sensitized to it. If the same input vector is shown to the network again after training, this node will provide maximum response. Experimentally it has been found that to achieve global ordering of the map the neighbourhood set around the winning node should initially be large so that it quickly produces a rough mapping. As more training sets are passed to the network, this neighbourhood set is reduced to force localized adaption of the network. The magnitude of the change decreases with time (where time means number of iterations for the training process) and with distance from the winning node. This is done by reducing the learning rate during the training.

Assume \( w_i (t) \) is the weight vector for the \( i \)th node at time \( t \), and \( X(j) \) is the input vector for the \( j \)th training case. On each step, a training case \( X(j) \) is selected and the winning node with smallest distance between weight vector and the input vector is selected
\[ \| X(j) - w_c(t) \| = \min_i \| X(j) - w_i(t) \| \]

Hence index \( c \) of the winning node is determined by

\[ c = \arg \min_i \| X(j) - w_i(t) \| \]

Update weights for node \( c \) and nodes within its neighborhood. The iterative adjustment of the weights based on this allocation of cases to nodes is referred to as the training of the SOM. The Kohonen SOM algorithm requires a kernel function \( K(i,c)(t) \) where \( K(i,i)(t) = 1 \), and \( K(i,c)(t) \) is usually a non-increasing function of the distance between nodes \( i \) and \( c \) in the grid space.

\( K(i,c)(t) = 0 \), for nodes that are far apart in the grid space (i.e., nodes which are not in the neighborhood). In the process as each training case is processed, the weight vector of the nodes is updated as indicated below:

\[ w_i(t+1) = w_i(t) + K(i,c)(t) L(t) \{ X(t) - w_i(t) \} \]

where \( K(i,c)(t) \) is the kernel function and \( L(t) \) is the learning rate which decreases during training. The kernel function \( K(i,c)(t) \) typically changes during training and decreases when the distance in the grid between node \( i \) and the winner node \( c \) becomes larger. The neighbourhood of a given weight vector is the set of nodes for which \( K(i,c)(t) > 0 \). A SOM works by smoothing the weights for nodes in neighbouring cells in the grid space. It is advisable to start with a large neighbourhood and gradually shrink the neighbourhood size during training, in order to avoid poor results. If interested in topological mapping, it is important not to let the neighbourhood size shrink all the way to zero during training. Indeed, the choice of the final neighbourhood size is the most important tuning parameter for SOM training. Determining a good neighbourhood size usually requires trial and error. In a Kohonen SOM, it is necessary to reduce the learning rate during training to obtain convergence.

Kohonen SOM is suitable for highly nonlinear multidimensional data. Because of its efficiency with very large datasets, we used SOM for clustering in the second study in this thesis. The selected SOM algorithm behaves as follows:
1. Number of clusters is specified as the products of the specified number of rows and columns. Each cluster corresponds to an output node in the map.

2. The initial weights for the output nodes are randomly selected using a random sample of observations.

3. The initial neighborhood size is set to half the size of the SOM grid. The neighbourhood size is gradually reduced to zero during the first 1,000 training steps.

4. Incremental training is used. The learning rate is initialized at $L = 0.9$ and linearly reduced to 0.02 during the first 1,000 training steps.

4.3.3 Time Series clustering:

Clustering time series has been a successful area of research in a wide range of fields. Clustering is an unsupervised learning task dividing the population into homogenous groups based on their behaviour. There have been many studies on different clustering algorithms for time series data. The choice of an appropriate clustering algorithm is dependent upon the characteristics of the data and the purpose of clustering (Liao 2005). Keogh, Lin and Truppel (2003) classified time series clustering broadly into two categories:

1. **Whole clustering:** in this approach similar time series from a set of individual time series are grouped into one cluster.

2. **Subsequence clustering:** sliding windows are extracted from a single time series and these windows are compared to find similarities and difference.

Two of the popular algorithms used for clustering and finding similarities between data points in raw time series data are: Hierarchical clustering and K-mean clustering based on Euclidean Distance. Euclidean distance is the most frequently used metric (Agrawal, Faloutsos, & Swami, 1993). Empirical comparison by Keogh and Kasetty (2002) has shown that Euclidean distance performs better than other distance measures when applied on the same datasets. However this method requires that the time series be of the same length.
In a survey study, Liao categorized the previous methods for clustering time series into 3 major categories according to the data used for clustering (Liao 2005):

1. Raw data based approach: clustering methods that are performed on the raw data (Golay, et al., 1998; Kakizawa, Shumway, & Taniguchi, 1998; Liao, et al., 2002). Each time series point represents a different variable. This means that all the series have to be of the same length.

2. Model-based approach: in this approach time series are described by a model such as an AR coefficient model (Maharaj, 2000); Markov chain model (Ramoni et al., 2001); ARIMA mixture model (Xiong & Yeung, 2002).

3. Feature-based approach: a set of features are extracted from the raw data and these features are used for clustering. This approach also allows series of differing length. The feature based approach has been used by several authors with cross-correlation function features by Goutte et al. (1999), perceptually important point features by Fu et al., (2001), time-frequency representation of the transient region features by Owsley et al. (1997).

Both the feature based approach and the model based approach are used in this research.

There are various methods which can be used to extract statistical features of time series such as the Discrete Wavelet Transform (DWT), Discrete Fourier Transform (DFT) and Compression based Dissimilarity Measures (CDM), but they all have a high computational complexity and require various conditions to be satisfied for the method to be successful. But even with more classical and statistical measures such as trend, seasonality, periodicity, serial correlation, skewness and kurtosis, useful sets of time series characteristic features can be extracted as measures.

Feature extraction can also be considered a means to obtain dimension reduction in a time series. Being able to extract the summarized characteristics of the time series, one can get a meaningful reduction in dimension so that long length datasets or different length time series can be reduced to a limited number of measures that are less sensitive to noise. These extracted features contain within themselves the distinct characteristic of the parent time series and hence, when utilized in a clustering algorithm as a finite set of
inputs, they can distinguish the similarity and the differences between the original time series (Wang, Smith, & Hyndman, 2006). Thus the main aim of the feature extraction is to produce a set of input measures containing the original characteristics of the parent time series which can be used in any of the clustering techniques (Wang, et al.). The approach proposed by Wang et al., aims to provide a method of clustering time series which is robust to missing data and variation in the lengths of the series. It follows the feature extraction principle for time series data but relies upon both statistical features and model parameter extraction.

In this thesis, we have also used a mixture of model based and feature based approaches for clustering. We have extracted some features from the time series and also from the AR (p) models fitted to the time series, and have used all these values as clustering inputs. The selected features used in our analysis are explained below.

Serial Correlation: The correlation between the successive values in the time series is called serial correlation. The Box-Pierce statistic can be used to evaluate the degree of serial correlation.

Skewness: Skewness is a measure to check the symmetry of the data. For a normal distribution, skewness is zero. Negative values for skewness tell us that the left tail is heavier than the right tail and positive values tell us that the right tail is heavier than the left tail.

Kurtosis: The measure of peakedness of data with a normal distribution having kurtosis equal to 3. A high value of kurtosis tells us that the distribution has a distinct peak near the mean, and then declines sharply with heavy tails, whereas a low value of kurtosis suggests a flat top near the mean.

Standard Deviation (STD): STD is a widely used measure of variability. It shows how much variation there is from the mean (average). A low standard deviation indicates that the data points tend to be very close to the mean, whereas a high standard deviation indicates that the data is spread out over a large range of values. STD is very sensitive to outliers.
Mean: The mean is the sum of all the time series values divided by the number of values present in the series. The mean is the most common measure of central tendency, but it is very sensitive to outliers.

We can now extract the basic measures of a time series by using the above defined global features and use this set of inputs to apply the appropriate clustering algorithms regardless of the length of the series or missing data.

This method of clustering time series, based on feature extraction measures is better than other clustering methods in the sense that it can cluster time series of varying lengths, which means the method is robust to missing data (Wang, et al., 2006). Unlike other alternative methods (such as k-means) this approach does not cluster the time series points based on a distance measure, rather it clusters global features extracted from individual time series, and therefore it can be applied on time series of different length. These features are then fed into any relevant clustering algorithm. Clustering based on global measures leads to more instinctive clustering results and is also robust to missing data (Wang, et al., 2006). Some of the advantages of this method discussed by Wang, et al., are as follows:

- When dealing with long time series (high dimensionality), some clustering algorithms become intractable. Feature extraction reduces dimensionality and therefore long length time series can be clustered very efficiently.
- Clustering algorithms based on a distance metric (i.e. Euclidean distance) cannot handle time series with missing data or time series of different lengths if actual points are used as inputs. However, this problem does not occur when a set of measures is extracted from the original time series.

In addition to feature extraction, we have incorporated the model-based approach in our analysis. In model based clustering, time series are described by model parameters and then clustering is performed on the estimated parameters. To choose an appropriate model, we looked at the customer’s daily spending pattern and noted a weekly pattern for most customers. This led us to fit an Autoregressive (AR) model with lag 7 to each time series, and 7 parameters were estimated.
The AR (p) model for an $X_t$ series is written as:

$$X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \varepsilon_t,$$

where $\varphi_1, \ldots, \varphi_p$ are the parameters of the model, $c$ is a constant and $\varepsilon_t$ is an error term.

The estimated parameters from the AR (7) model, along with the global characteristics extracted from the time series explained in the previous section were used as input for clustering. Both of the above methods (feature extraction and model based approach) create new variables from the original time series which represent the monthly transactional behaviour of the customers. These variables were proven to be significant in the prediction of delinquency in later periods.

Our attention now moves from exploratory approaches to classification and prediction. Classification techniques have been widely used in credit scoring. Due to simplicity and robustness, these techniques are still commonly used by most financial institutions for building scorecards. Classification techniques are appropriate when the dependent variable is categorical. In credit scoring the dependent variable is a categorical variable classifying the customers into two categories “good risk” or “bad risk”. Conventional/traditional statistical methods used in the score cards industry are Linear discriminant analysis (DA), Logistic Regression and Decision Trees. Typical terminology used in the industry refers to predictor variables as “characteristics” and the values that these variables take are referred to as “attributes”. This terminology is used throughout this thesis.

### 4.4 Linear discriminant analysis (LDA)

Linear discriminant analysis (DA) is the most popular method for building credit scoring models in the literature. DA is used to classify observations into two or more mutually exclusive groups by using the information provided by a set of characteristics.
DA was one of the first statistical techniques used for discriminating distinct populations.
This technique is referred to as two-group linear discriminant analysis when the observations are in only two groups. When there are more than two groups, this technique is called multiple linear discriminant analysis. However, in this thesis, while only two-group linear discriminant analysis is used, the term Discriminate Analysis (DA) is used for the sake of brevity.

Linear discriminant analysis was first introduced by Fisher in 1939 for multi-normally distributed data. In two group Linear discriminant analysis, a linear discriminant function ‘y’ that passes through the centroids of the two groups can be used to discriminate between the two groups.

To get a linear discriminant function, each independent attribute is multiplied by its corresponding weight before being added to the constant term with the weights chosen so as to maximize the discrimination between groups.

The linear discriminant function y is calculated for each customer using the formula:

\[ y = c + \sum w_i x_i \]

where y is the discriminant score, c is a constant; \( w_i \) is the weight of the \( i^{th} \) characteristic and \( x_i \) is the value of the \( i^{th} \) characteristic. The final decision function for a linear discriminant analysis is determined by a cut-off score. The cut-off score forms the dividing point, which is used to classify observations into groups based on their discriminant score values. For a 2 group LDA the cut-off score is calculated by averaging the group means or centroids, which are in turn obtained by averaging the linear discriminant function of all the observations within a particular group (Hair, Anderson, Tatham, & Black, 2006).

Linear discriminant analysis assuming multi-normally distributed data is commonly used for classification problems due to its simplicity and availability in statistical software. Despite receiving widespread attention in the credit scoring literature, DA has its limitations. For example categorical independent characteristics cannot be handled
with DA and DA assumes that the predictor variables are normally distributed\textsuperscript{8} with a linear and a homoscedastic relationship. A common procedure to achieve normality is the use of transformations such as standard (base 10) log or natural log to transform data before estimating the discriminant function. However, Eisenbeis (1977) points out that the application of a transformation may change the interrelationship between the variables; it may affect the relative position of the observations in the group and for positively skewed distributions it means that values which are larger get less weight when transformed than smaller values. Another limitation is DA’s dependence on a relatively similar sample size for the groups and an assumption of equal variance-covariance matrices for each group in the population. Reichert et al., (1983) demonstrated that the assumption of equal covariance matrices is unlikely to hold in practice because there are categorical variables coded using binary variables for which the variance depends on the mean. Violation of this assumption affects the significance test for the differences in group means and the suitable form of the classification rule. Note that this assumption is made only in terms of linear discriminant analysis which is the only form of DA used in this research (If the dispersions of the groups are unequal, then quadratic procedures are appropriate (Eisenbeis & Avery, 1972). However, Márquez (2008) pointed out that “the added uncertainty in estimating two covariance matrices instead of one makes the “quadratic” decision rule less robust than the linear decision rule, and the slightly better accuracy is not worth the effort in general”. Mark and Dunn (1974) suggested that when dealing with small sample sizes and a large number of variables, Linear discriminant analysis gives more efficient estimates of the expected error rates (compared to quadratic DA) even if the population dispersions are unequal. For the above reasons, the use of linear discriminant analysis for credit scoring has often been criticized (West, 2000). Nevertheless, some authors have found that the predictive performance of DA in credit scoring is superior to that of artificial neural networks, genetic algorithms and decision trees (Yobas, Crook, & Ross, 2000).

\textsuperscript{8} However, Reichert et al. (1983) conclude that non normality of significant portion of credit information is not a critical limitation.
4.5 Logistic Regression

Logistic regression (LR) has emerged as the most suitable classification technique when
the dependent attribute is binary and the independent attributes are continuous,
categorical or both. Since the outcome of credit scoring is usually binary, logistic
regression is probably the most suitable classification approach for credit scoring.

In fact, based on our discussion with several credit analysts at a leading Australian bank,
logistic regression is the main technique used by most financial institutions for credit
scoring. It is designed to predict the probability (p) of future bad customer (Y=1). The
odds for this event can be predicted as a linear function of a customer’s attributes (x_i,
i=1, 2…k).

The binary logistic regression model has the form:

\[
\ln\left(\frac{p}{1-p}\right) = \beta_0 + \sum_{i=1}^{k} \beta_i x_i
\]

where p is the probability of the event of interest (in our case it is the probability of a
future bad customer), \( \beta_0 \) is the constant; \( \beta_i \) is the coefficient for the \( i_{th} \) characteristic and
\( x_i \) is the \( i_{th} \) characteristic (independent variable). \( \ln(p/(1-p)) \) is known as the log odds
and is sometimes referred to as the logit transformation. Suppose we have n customers
\( (j=1,2, ..., n) \) who are assumed to be statistically independent. The inverse
transformation converts the log odds back to probabilities for bad future customers
\( (Y=1) \) and good future customers \( (Y=0) \) as shown below:

\[
p = P(Y = 1) = \frac{\exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}{1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)} = \frac{1}{1 + \exp(-\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}
\]

\[
1 - p = P(Y = 0) = \frac{1}{1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i x_i)}
\]
The outcome of logistic regression is the estimated probability of the event which must have a value between 0 and 1. A cut-off point for this probability, for example 0.5, is needed to classify an applicant as either ‘future bad’ or ‘future good’ customer.

The coefficients and constant term are estimated using Maximum likelihood estimation (MLE). Let $p_j$ be the probability of $(Y_j=1 ; j=1,2,\ldots,n)$. This is the probability of the jth customer becoming a future bad customer. And let $(1- p_j )$ be the probability of $(Y_j=0 ; j=1,2,\ldots,n)$. This is the probability of a future good customer. The likelihood function expresses the probability of observing the data as a function of the unknown parameters. The likelihood of observing the values of $Y$ for all the observations is given as:

$$L = p(Y_1,Y_2,\ldots,Y_n)$$

Since we assume that observations (customers) are independent, the overall probability of observing all the $Y_j$’s can be written as the product of the individual probabilities:

$$L = \prod_{j=1}^{n} p(Y_j | X_{j1}, \ldots, X_{jk}) = \prod_{j=1}^{n} p_j^{y_j} (1 - p_j)^{1-y_j}$$

$$L = \prod_{j=1}^{n} \left\{ \frac{\exp(\beta_0 + \sum_{i=1}^{k} \beta_i X_{ji})^{y_j}}{1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i X_{ji})} \right\}^{y_j} \left\{ \frac{1}{1 + \exp(\beta_0 + \sum_{i=1}^{k} \beta_i X_{ji})} \right\}^{1-y_j}$$

where $X_{j1}, \ldots,X_{jk}$ are the values of independent variables for the jth customer based on sample of n customers.

Taking log on both sides and doing a little algebra we get:

$$\ln(L) = \sum_{j=1}^{n} \beta'y_jy_j - \ln(1 + \exp(\beta_0 + \beta'x_j))$$
where $\beta = (\beta_0, \beta_1, \ldots, \beta_k)'$ and $x_j = (1 \ x_{j1} \ x_{j2} \ldots x_{jk})'$ is the vector of independent variables. Because logarithm is an increasing function, whatever maximizes the logarithm will also maximize the original function (Allison, 1999). We have to select the $\beta$ values which maximize the above equation and the resulting values are known as maximum likelihood estimators (MLE’s). This is done by the methods of calculus. One well known method is to find the derivative of the function with respect to $\beta$, set the derivative to 0 and solve for $\beta$. The maximization process is complicated compared to least square estimation used in LDA, requiring an iterative approach. However this iterative solution procedure is available in most popular statistical procedures such as SAS.

The advantage of logistic regression over other statistical models is that it can handle categorical variables as predictors as well as metric variables, and as such does not involve some of the assumptions necessary for linear discriminant analysis (i.e. normal distribution, linear and homoscedastic relationships between the dependent and independent variables). Harrell and Lee (1985) found that even when all the assumptions of LDA are satisfied, LR is as efficient and accurate as LDA. However, although logistic regression has fewer assumptions compared to DA, it has been found that logistic regression is as good a classifier as linear discriminant analysis but perhaps no better (Harrell & Lee).

### 4.6 Decision trees

Decision trees are flexible in terms of the types of variables they can incorporate as predictors and are particularly useful for exploratory analysis. A tree is a hierarchy of questions about the observations, and the final classification made depends on the answers to these questions. The tree consists of a series of nodes which split the observations into descendent subsets. This method works sequentially and it is often referred as “recursive partitioning”. Each descendent node is purer in terms of class separation than the parent node.
There are several reasons why this method is appropriate for credit scoring. They are as follows (Hand and Henley 1993):

1. Decision trees are non-parametric classification rules which are able to fit complex non-linear relationships. Credit scoring datasets are multidimensional and the presence of non-linear relationships is almost certain. Since decision trees do not have a restrictive parametric form, this method is appropriate for credit scoring.

2. Decision trees construct local regions based upon the nature of the response variation and so they do not suffer from data sparsity in high dimensions, making efficient use of the available data.

3. Decision trees consist of a series of sequential nodes and each node is purer than its parent node in that there is less outcome diversity. The underlying decision process can be better represented by a sequential process rather than a model which considers all the characteristics simultaneously.

4. Decision trees are easy to monitor and understand (Makowski, 1985) and are easily represented using interpretable logic statements rules.

5. A decision tree can use the values of an input surrogate variable as back-up when missing data prohibits the application of a splitting rule.

Decision trees have been extensively used to develop credit scoring models (Carter & Catlett, 1987; Makowski, 1985; Mehta, 1968). Good classification performance is achieved in several of these studies (Boyle, et al., 1992; Srinivasan & Kim, 1987). For this reason we use decision tree as one of the selected methods in this research. One problem faced when using decision trees is that the sample size available for modeling at each split is cut drastically and this can affect the performance of the model. Another disadvantage of this method is that variable selection in a classification tree is biased in favor of variables having more attributes (Breiman, et al., 1984).
An empirical tree represents a segmentation of the data that is created by applying a series of simple rules which, based on the value of a single characteristic, assigns each observation to a node or segment. Several rules are applied one after another resulting in a hierarchy of segments within segments. The original segment is called a root node and it contains all the data points. The final or terminal nodes are called the leaves. A classification decision is made for each leaf and is applied to all the data points in that leaf. The type of decision depends on the context. In credit score predictive modeling, the decision is simply the predicted value (good/bad).

There are two phases in constructing a decision tree, a growing phase and a pruning phase. The selection of rules to split the node is called the growing phase and the decision when to declare a node as a terminal node is known as the pruning phase. The decision tree starts with one node that contains all the points in the sample under study. The growing phase involves repeatedly splitting nodes into better class separation nodes, considering all the characteristics and finding the best possible splitting point for each characteristic. The tree is grown using the best splitting rules until it is not possible to improve class separation using additional splits or the number of observation at a particular node is less than a specified threshold.

There are a number of splitting criteria for different target measurement levels. In this research, the target variable is a binary measure (good, bad). There were three splitting criteria in the software used for the analysis (SAS E-Miner) for a binary or nominal target variable. These splitting criteria use chi-square tests, gini reduction and Entropy reduction. All three criteria were tested and the best splitting criteria (based on ROC curve and Gini coefficient) was selected.

Initially the maximum number of branches from a node and the minimum number of observations in a terminal node or leaf are set by the analyst. The minimum number of observations in a leaf prevents the splitting of nodes that have only a few observations. The CHi-squared Automatic Interaction Detector (CHAID) developed by Hartigan in 1975 (Hartigan, 1975), is the most widely used method for splitting the tree nodes especially because it is distributed as part of the popular statistical packages SAS and
SPSS. Consider a categorical predictor with $n$ categories and a target variable with $z$ categories then

$$\chi^2 = \sum_{i=1}^{n} \sum_{j=1}^{z} \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

$\chi^2$ is Pearson's cumulative test statistic, which asymptotically approaches a $\chi^2$ distribution. $O_{ij}$ is the observed frequency for cell $(i,j)$ and $E_{ij}$ is the expected frequency for cell$(i, j)$ if the null hypothesis of independence is true.

The $p$-value for this test statistic is based on the chi-square probability distribution with $(n-1)$ degrees of freedom and is generally estimated using an algorithm. The $p$-value represents the probability that the chi-square test statistic is as extreme as or more extreme than observed, if the null hypothesis of independence is true. There is a different chi-square distribution for each value of $n$.

The advantage of this method is that it stops growing the tree before over-fitting occurs. This method is restricted to categorical predictor variables; therefore continuous variables need to be redefined into classes. If this method is to be applied with continuous predictors, the splitting criteria for this method are based on the $-\log (p$-value) measure sometimes called the LOG WORTH. For binary splits on binary or interval predictors, the optimal split can always be found. Rather than over-fitting the data and then pruning the tree back this method stops growing the tree before over-fitting occurs.

The Gini index (CART) is one of the most popular methods for developing trees and was first published by Breiman et al., (1984). This method builds a binary target tree by splitting the observations at each node according to the following Gini index (Berry & Linoff, 2000). The higher this index the greater the node impurity. The Gini index for a node is defined as

$$I(\text{node}) = 1 - (p_1^2 + p_2^2)$$
$$= 2p_1 (1 - p_1)$$
where \( p_1 \) is the proportion of bad risk customers in a node and \( p_2 \) is the proportion of good risk customers in a node. A pure node has a Gini of zero and a node with equal number for both classes has a Gini of 0.5. When there are \( k \) classes with relative frequency \( p_i; i=1, 2, \ldots, k \). Gini looks for the largest class and strives to isolate it from all others. The best variable and the best split point to achieve this, is found by comparing the effects on the Gini index. The goal is to maximize the reduction in the Gini index. Once the first split is made, Gini continues attempting to split the data that requires further segmentation, using the same strategy. Gini attempts to separate classes by focusing on one class at a time. With this method over-fitting tends to occur making it necessary for pruning to occur before the tree can be used for classification purposes.

*Entropy* is another measure of node impurity. For a binary target, Entropy is defined as

\[
I(\text{node}) = - p_1 \log p_1 - p_2 \log p_2
\]

A node which has smaller entropy than another node is more homogenous. A pure node has entropy equal to zero and a node which has an equal proportion of responders and non responders has entropy of 1. The choice of splitting rules for all the above trees are described below.

As the tree is constructed for each node, a number of splitting rules are evaluated. A statistic that measures the strength of the rule is computed for each input and split point, i.e., the tree searches for the split with maximum worth or log worth, subject to the limit on the number of branches and the limit on the minimum number of observations assigned to a branch. The measure of worth tells us how well a variable divides the data into each target class. If the Entropy or Gini reduction criterion is selected, then the computed statistic is WORTH. WORTH measures the reduction in variance for the split. If the Chi-square criterion is selected, then the computed statistic is the LOGWORTH. The splitting rules are sorted by LOGWORTH or WORTH depending on the selected splitting criteria. For both LOGWORTH and WORTH, larger values are better because they indicate a large reduction in node impurity.
The Entropy and Gini reduction criteria measure worth as follows when a node splits into branches $b=1, 2…$

$$ \text{WORTH} = I(\text{node}) - \sum \text{P}(b) \cdot I(b) $$

where $\text{P}(b)$ is the proportion of observations in the node assigned to branch $b$ and $I(b)$ represents the Entropy or Gini in the node branches.

The Chi-square test uses LOGWORTH;

$$ \text{LOGWORTH} = -\log (p\text{-value from Chi-square}). $$

For these criteria, the best split is the one with the smallest $p$-value. By default, the $p$-values are adjusted to take into account multiple testing. Unlike Chi-Square trees, Gini and Entropy trees tend to over-fit the training data, making it necessary for these trees to be pruned back in order to improve their classification accuracy. This is done by comparing the error rated of earlier trees with the final tree using a fresh set of validation data as shown in Figure 4.2. The tree size with the lowest error rate is chosen.

![Figure 4.2 Pruning back to get optimum tree size](image-url)
4.7 Artificial Neural Networks

Artificial Neural Networks represent one kind of artificial intelligence system. Artificial intelligence systems are information processing and modeling systems which mimic the learning ability of biological systems in understanding an unknown behaviour. These systems can be classified into brain-nervous systems, genetic systems and immune systems.

Intelligent systems techniques have been used for understanding financial problems such as credit scoring. Goonatilake and Treleaven (1995) attribute the remarkable success of the intelligent systems in solving financial problems to five main features. These features are:

1. Learning: Intelligent systems have the ability to learn decisions and tasks from a large historical database. This is the most important feature of intelligent systems.
2. Adaptation: Intelligent systems have the ability to adapt to changes in the environment such as policy changes, new regulations or changes in economic conditions.
3. Flexibility: Intelligent systems have the ability to make decisions and perform tasks even in the presence of incomplete and unreliable data.
4. Explanation: Intelligent systems have the ability to explain how the decisions were reached. This is useful in situations (legal or organizational reasons) where it is compulsory to provide explanations of how decisions have been made.
5. Discovery: Intelligent systems have the ability to discover new relationships that were previously unknown.

However not all intelligent system techniques exhibit all of these five features. Each technique has its own strengths and weaknesses and thus cannot be used as a general-purpose tool for every type of problem.

Artificial neural networks (ANNs) and genetic algorithms (GA) are good examples of successful applications of biological metaphors to devise solutions to computational problems. There has been a growing interest in the fields of artificial neural networks
and genetic algorithms in the past few decades. Application of ANN increased in 1990s with an improvement in our understanding of how the mind and the brain work. A 1997 report by the Stamford, Connecticut–based Gartner Group described these intelligent systems as being at the top of the five key technologies that will have a great impact across a wide range of industries within the next five to ten years (Metaxiotis & Psarras, 2004).

According to Chatterjee (2000), ANNs can take the place of traditional statistical approaches in many financial decision making instances.

The main advantage of artificial neural networks is that, with a suitable number of hidden nodes they can approximate any nonlinear function to an arbitrary degree of accuracy (Hornik, et al., 1989) and provide reliable ways of obtaining solutions to a variety of problems that often cannot be dealt with using traditional methods (Goonatilake & Treleaven, 1995).

Many studies have shown that ANNs have outperformed traditional statistical modeling (Desai, et al., 1996; Lancher, et al., 1995; Malhotra & Malhotra, 2003; Sharda & Wilson, 1996). Hence ANNs were selected as one of the models used in this research to describe customer behaviour.

Artificial Neural Networks (ANNs), also often known as Neural Networks (NNs) are inspired by the functionality of the nerve cells in the brain. As said by Parker (2006) ANNs are based on the biological concept of learning and memory creation relying on the construction of interconnected nodes. ANNs consist of a set of connected input/output units where each connection has weights associated with it. During the learning phase, the network learns by adjusting the weights so as to be able to predict the correct class label for each input sample.

While ANNs can achieve high predictive accuracy, they have been criticized for their poor interpretability. It is difficult to understand and interpret the symbolic meaning behind the learned weights. Also ANNs require a long training time. These factors have limited the application of NN in the field of credit scoring (Chung & Gray, 1999; Craven & Shavlik, 1997). This is why financial institutions still choose the traditional
classification modeling over ANNs even though many studies have shown that ANNs perform significantly better than conventional classification techniques such as linear discriminant analysis and logistic regression analysis (Desai, et al., 1996; Lee & Chen, 2005; Malhotra & Malhotra, 2003). Further research is required for the interpretation of ANN models.

There are numerous types of ANN. The simplest form of ANN is a perceptron having a single input and single output, equivalent to simple linear regression or having multiple inputs and a single output which is effectively multiple linear regressions. The most common ANN is a multi-layer perceptron (MLP) which uses feed-forward architecture. In such networks all the signals flow in one direction. There are also network patterns which allow feedback paths, but such loops may introduce computational complexity and problems with local minima.

Complex statistical relationships can be modeled by choosing a suitable MLP network architecture. This network is known as a supervised network because it requires an output in order to learn. The goal of this type of network is to create a model that correctly maps the input to the output using historical data, so that the model can be used in future to predict the attributes of dependant variables.

MLP usually consists of an interconnected group of nodes arranged in 3 layers, namely the input layer, the output layer and a layer between the two, called the hidden layer. MLP is a feed-forward neural network, i.e. the information moves in only one direction, forward. Figure 4.3 shows a three layer feed-forward network.

![Feed-forward three layer ANN](image-url)
The input layer consists of input nodes that represent each of the predictor characteristics. The output layer is the last layer in the network and represents the classification decision. The hidden layer consists of hidden nodes which assist the propagation of feed-forward information from the input layer to the output layer (Behara, Fisher, & Lemmink, 2002). An MLP can have any number of inputs, at least one hidden layer with any number of units (nodes) and any number of outputs, with one node for each output attribute. The nodes in different layers of the network are connected by links with variable weights, which correspond to the parameters of statistical models. Stimulation is applied to the inputs of the first layer. Each node (neuron) receives data from the node in the previous layer and the signals are multiplied by a separate weight. The weighted inputs are combined and passed through a limiting function. The limiting function scales the output to a fixed range of non-linearly scaled values between 0 and +1 or between -1 and +1, and then the output of the limiter is transmitted to all the nodes in the next layer. This process is shown in Figure 4.4.

**Figure 4.4** Structure of a node with a sigmoid function as limiter.

Figure 4.4 illustrates the activation function at each hidden or output node. The activation function consists of a combination function ($\sum$) and a limiting (or transfer) function ($\Phi$). A combination function combines the values received from preceding nodes into a single number called the net input. In MLP both output and hidden layers use linear combination functions. A linear combination function computes a linear combination of the weighted values feeding into the node, and then adds the bias value
which acts as an intercept. This net input at each of these nodes is then transformed by the limiting function.

A Sigmoid limiting function, also known as logistic functions, produce values between 0 and 1 (Warner & Misra, 1996), and is given by:

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

where $f(x)$ is strictly positive and defined for all values of $x$. Various other limiting functions are shown in Figure 4.5.

<table>
<thead>
<tr>
<th>Name</th>
<th>Function $y=f(x)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Logistic</td>
<td>$\frac{1}{1+e^{-x}}$</td>
</tr>
<tr>
<td>Tanh</td>
<td>$\tanh(x)$</td>
</tr>
<tr>
<td>Gaussian</td>
<td>$e^{-x^2/2}$</td>
</tr>
<tr>
<td>Linear</td>
<td>$x$</td>
</tr>
</tbody>
</table>

*Figure 4.5  Different types of activation function.*

MLP commonly uses sigmoidal limiting functions in the hidden layers and a linear function in the output layer.
Figure 4.6  An example of MLP with one hidden layer

The hidden layer makes the network more powerful by enabling the network to recognize more patterns. Increasing the size of the hidden layer increases the power but introduces the risk of over-fitting. Choosing the right number of hidden layers is a crucial step. We have to choose that number which minimizes the misclassification rate. In practice, usually, only one hidden layer is preferred for prediction, since a single hidden layer network is sufficient to model any complex system (Hornik, et al., 1989; Zhang, Patuwo, & Hu, 1998) Therefore in this thesis the designed network will have only one hidden layer.

Although one hidden layer is usually sufficient, there is no theory to tell you how many hidden units are needed to approximate any given function. Trial and error can be applied to get the number of hidden units which minimizes the misclassification rate. Sometimes networks with two or more hidden layers may require fewer hidden units and weights than networks with one hidden layer. We have used the trial and error approach with the range from 1 to 50 nodes to determine the appropriate number of hidden nodes for the hidden layer in our networks.

MLPs are known as universal approximators because, with just one hidden layer, they can learn to approximate virtually any function to any degree of accuracy. This assumes that there is enough data, enough hidden nodes, and enough training time.
The network is defined by the weights on the paths between the nodes. Construction of the network involves a learning algorithm to estimate these weights. For MLP, most learning algorithm uses a “backpropogation rule” (Rumelhart & McClelland, 1986). This learning algorithm modifies the network weights in order to minimize the difference between the estimated and actual outputs. A backpropogation network learns by example. One of the example cases is applied to the network, and the network produces some output based on the current state of its weights. This output is compared to the actual output. A mean-squared error signal is calculated for the expected and actual output. The error value is then propagated backwards through the network, and small changes are made to the weights in each layer. The weight changes are chosen so as to reduce the error signal for the case in question. This process is repeated for each of the example cases, then back to the first case again, and so on. The cycle is repeated until the overall error value drops below some pre-determined threshold. At this point we say that the network has learnt the problem "well enough". Although using such an algorithm has a risk of not converging to a global minimum, it has been shown to be very efficient (Mitchell, 1997).

Note that in the case of a Multi Layer Perceptron, we use the training dataset to find the “optimal” weights with the back-propagation rule, but we use the validation dataset to find the optimal number of hidden nodes and to determine the optimum number of cycles for the back propagation algorithm. We use the test data to estimate the prediction/classification error rate after we have chosen the final model.

4.8 Two-stage model:

The previous classification models were used in this study to differentiate between high risk and low risk customers. However it is also helpful to have information on the amount owed by the delinquent customers along with their delinquency status in order to recognize high risk customers. For this reason we have used two stage models to predict the outstanding amount and delinquency status of each customer.
In the two stage model, prediction is done in two stages. In the first stage, the class target variable (delinquency status) which is a categorical is predicted and at the second stage the interval target (outstanding amount) is predicted. Different types of model can be used for these two purposes. The model used for predicting the class target is known as a class model and the model used to predict the interval target is known as a value model.

The structure of the two stage model is as follows:

1. **Class model** specifies the model that is used to fit a model for the categorical target variable in the first stage. The available class models in SAS Enterprise Miner for two-stage models are Trees, Logistic Regressions, Multi Layer Perceptrons (MLP), Radial Basic Function (RBF) and Generalized Linear Model (GLIM). Selecting the right model depends on the data and the outcome. The analyst should decide the most appropriate model for the data. In this research we have compared the performance of trees, logistic regression and MLP models. These methods are explained in previous sections. The model with minimum misclassification rate for the test data is selected as the best model.

2. **Value Model** specifies the model that is used to fit a value model for the interval target variable in the second stage. The available models are Tree, Regression, MLP, RBF and GLIM. For predicting the interval target we have applied tree, regression and MLP.

3. **A transfer function** determines how the class target variable is incorporated into the value model of the second stage. There are two available options knows as probability and classification transfer functions. The *probability* transfer function uses the posterior probability of the class event as an input variable to fit the value model. The *classification* transfer function uses the predicted classification as the input. We have used the probability transfer function option because the estimated probabilities provide a more accurate reflection of risk.

4. **Filter** specifies how observations of the training datasets are excluded in training the value model. Since we wanted to consider all observations, we have used no filter in our analysis.
4.9 Summary

In this chapter we have introduced the clustering and Markov Chain methodology which has been used in this thesis. In addition the methods used to develop models for predicting future bad customers have been explained in detail. In the next chapter we use some of these methods for the analysis of the Study 1 data.
Chapter 5

EXPLORATION OF CUSTOMERS’ BEHAVIOUR FROM A MARKETING PERSPECTIVE

It is common sense to take a method and try it. If it fails, admit it frankly and try another. But above all, try something.
- 32nd President of United State, Franklin D. Roosevelt
Table 5.1 *Outline of chapter five*

| 5.1 | Introduction |
| 5.2 | Aim of the Pilot study |
| 5.3 | Data description for Pilot study |
| 5.4 | Analysis |
| 5.4.1 | Study 1 (clustering based on transaction amount) |
| 5.4.2 | Study 2 (clustering based on merchant category code) |
| 5.5 | Discussion |

5.1 **Introduction:**

Getting access to banking data is not easy for the reason of the confidentiality issues that banks face. The main dataset for this research comes from a leading Australian bank. It is a very large credit card dataset including 1.5 million customers’ transaction details. While waiting for the main dataset a small dataset was analysed to get more understanding of customer behaviour and to trial methods for the main study. This chapter explains in detail the analyses conducted in this pilot study. The outline of this chapter is shown in Table 5.1 above.

5.2 **Aim of the Pilot study**

The intended outcome of this research is a greater understanding of the evolution of customer behaviour. It is expected that this knowledge will enable the bank to optimize the customer’s banking experience while minimizing banking risk and maximizing profit.

The credit card data for this chapter comes from a second Australian bank. This dataset was limited to only a few variables. Transaction amounts and the Merchant names and codes were provided to us for a random sample of customers. Based on what was available we applied different approaches gaining better understanding of the methodology to be used in the follow up study. In this chapter we aim to identify
customers who have similar spending behaviour. In addition we develop a methodology for monitoring changes in customer behaviour that allows early detection of a change in a customer’s financial circumstances. Also in this study we consider changes in the type of expenditure. This is an important aspect of research for the banking industry with implications for their marketing strategies. For example, banks could come up with loyalty programs. Using this transaction amounts and what type of product the customer uses, banks could sell new products to them and reward them if they regularly use their credit cards.

Clustering is a good way of beginning to understand individual behaviour. In this thesis we wish to cluster credit card customers into groups of customers with similar behaviour and based on the characteristics of each cluster, predict the future behaviour of customers.

The main contribution of this study is the selection of new methodology and the choice of new variables for monitoring financial behaviour. The clustering of daily transaction time series, based on global characteristics extracted from these time series rather than using the whole time series directly, is used in both our studies: The resulting clusters allow the comparison of different groups of customers for marketing purposes and also for identifying potentially risky customers.

Another contribution of this research involves the use of product based input variables for examining financial behaviour. We have not only used the amount of transaction as an input but also divided the transaction amounts based on the types of product bought. Observing where the customer spends their money tells us about the financial circumstances of the customers. For example, if a customer spends money on holidays and sports, he is probably more financially sound than a customer who only spends money on petrol and food.

5.3 Data description for Pilot study:

The data used for the pilot study includes transaction amounts over a 6-month period from June 1, 2006 to November 30, 2006 for 1000 customers. The variables in the dataset which are referred to for this chapter are as follows:
• *Id* is the customer’s id,
• *Amount* is the total amount the customer spends per transaction.
• *Date* is the date when the transaction takes place.
• *Merchant category name* is the type of merchant where the transaction takes place.
• *Merchant category code* (MCC) is the code given to each merchant category by the bank.

This data was not complete; it only gave us information on transaction amount. There was no information on the payment amount or the status of the customer in terms of delinquency. Based on what was provided to us we aimed to produce some meaningful clusters which could identify the customers with similar behaviour.

### 5.4 Analysis

Banking data are very large in size. This makes it difficult to monitor every transaction for each customer. Clustering is a good way to monitor individual behaviour with such large datasets.

Clustering is the process of organizing objects into groups whose members are similar in some way. In this thesis we aim to cluster credit card customers based on credit card usage behaviour.

Clustering can be performed directly on the raw data, but in our analysis we chose a more sophisticated method of clustering allowing a more useful interpretation of the clusters. We decided to aggregate the transaction amount to daily level to form a daily time series for each customer. A model was fitted to each customer’s time series and time series parameters were extracted. Also some other features (such as moments) were extracted from these time series. Clustering was then performed on the variables extracted from the time series for each customer.
There is no absolute "best" clustering criterion, which is independent of the final aim of the clustering. Consequently, it is the user who must supply this criterion, in such a way that the results of the clustering will suit their needs.

Ward's method is a minimum distance hierarchical method, which considers the union of every possible cluster pair at each step. The cluster pair which minimizes the increase in error sum of squares, defined as the sum of squared deviations from the combined cluster centroid, is selected. In this chapter we chose hierarchical clustering using Ward’s method because the dendrogram provides a useful picture of the clustering when sample sizes are relatively small.

The analysis for this chapter is performed in two separate studies. In the first study, customers are clustered by considering their daily transaction amount. In the second study customers are clustered based on how much they spend for each MCC code. These studies are explained in the next section.

5.4.1 Study 1 (clustering based on transaction amount):

This study aims to cluster customers based on transaction amount. Study one consists of two parts. In the first part (part a) 6 months of daily transactions for each customer are considered. For the second part (part b) we separate each month of transactions for each customer, clustering the customers in each month and then checking how the customers move from one cluster to another over the 6 month period. To start the analysis we need to extract the desired information from the data given to us.

5.4.1.1. Preparation of data:

For the first study the variables used are transaction amount, date and id. Amounts and dates were provided to us only if there was a transaction on a particular day. If there was no transaction on any day, that date did not appear in the data for that particular customer. The transactions were aggregated to daily level for each customer. If there was no transaction for a particular customer on a particular day, the transaction amount was zero; if there was more than one transaction per day, for example 3 transactions on
the same day, then three separate transactions were given for that date. We created a new variable called *transaction amount* which aggregated the transactions to daily level. *Transaction amount* is the amount spent per day and not per transaction. For example if there were two (or more) transactions on a particular day from a particular customer then the *daily amount* would be the sum of both (or all) transaction amounts for that particular day.

A new binary variable called *presence of transaction* was also created. If a customer spends any amount on a particular day, it implied that there has been a transaction on that particular day, therefore presence of transaction would be “1”, and if a customer does not use his credit card on a particular day then presence of transaction would be “0”. Therefore the final daily data includes 4 variables (id, transaction amount, date, presence of transaction) for 1000 customers and for each customer we have 6 months of data starting from the start of June to the end of November.

*Study 1 - Part a) 6-month time series:*

This is a retrospective study looking over the last six months to cluster customers aiming to recognize any transaction patterns based on customers’ daily transaction amount and presence of transactions over the 6 months. For each customer daily transaction amount and presence of transaction for 6 months are considered as two time series. Each of these time series includes 183 days.

Analysis starts with fitting of an AR (p) model to each of the 2 time series for each customer. This was done using “automatic ARIMA model fit” in SAS. Looking at the expenditure (transaction amount), a weekly pattern was common for most of the customers. Therefore, an AR (7) model was fitted to all the series including transaction amount and seven AR parameters were estimated. In addition, mean, standard deviation, kurtosis and skewness were calculated for each time series on transaction amount. Clustering based only on the AR (7) parameters from the daily transaction amount series produced only a few clusters with the majority of the customers (89%) classified into one cluster. This was deemed to provide insufficient information for monitoring behaviour.
Time series which included the binary sequence of presence or absence of transaction over 183 days were then analysed in the same manner to produce 7 Autoregressive parameters and 4 moments (mean standard deviation, skewness and kurtosis). The extracted parameters were merged to form the input dataset. This input data includes 24 variables for each customer (7 autoregressive parameters and 4 moments from transaction time series, 7 autoregressive parameters and 4 moments from presence of transaction time series, date and id).

Standardization was performed on all the variables since the variables were measured using different scales. Using the 22 variables (excluding id and date from input data), customers were classified into different clusters, using an agglomerative hierarchical clustering with distances measured using Ward’s method. The jump in the Cubic Clustering Criterion (CCC) shown in Figure 5.1 suggested 8 and 57 clusters. 57 clusters included many outliers (most of the 57 clusters contained only one customer). The pseudo F and pseudo-t squared statistic (a transformation of pseudo F) shown in Figure 5.2 suggested that the data should be divided into 8 clusters. The pseudo $t^2$ statistic can be applied only to hierarchical methods. A small value of the pseudo $t^2$ statistic and a larger pseudo $t^2$ for the next cluster fusion suggest the appropriate number of cluster (Cooper and Milligan 1988).

![Figure 5.1 Cubic clustering criterion](image)
Having too many clusters including only one customer does not help us in grouping customers with similar pattern; therefore we divided the customers into 8 clusters. Looking at the number of customers in each cluster, we can easily identify the outlier clusters. Maximum frequencies were seen in cluster 5 and 4 with 341 and 311 customers respectively.

Table 5.2 below shows the mean and standard deviation for daily transaction amount and the binary daily variable for transaction presence for each cluster.

Table 5.3 shows the characteristics of each cluster in terms of the first and seventh AR parameters ($\Phi_1$ and $\Phi_7$) for the above 2 series: High values for the $\Phi_7$ parameter suggest weekly shopping while high values for the $\Phi_1$ parameter suggest more irregular shopping behaviour, even daily shopping.
### Table 5.2  Means and Standard Deviations for Clusters

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>N</th>
<th>Daily Presence of transactions</th>
<th>Daily Transaction amount</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Mean</td>
<td>Standard deviation</td>
<td>Mean</td>
<td>Standard deviation</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>253</td>
<td>0.048</td>
<td>0.039</td>
<td>11.79</td>
<td>13.95</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>50</td>
<td>0.005</td>
<td>0.002</td>
<td>11.43</td>
<td>30.27</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>4</td>
<td>0.019</td>
<td>0.016</td>
<td>1.64</td>
<td>1.63</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>311</td>
<td>0.397</td>
<td>0.208</td>
<td>67.62</td>
<td>47.11</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>341</td>
<td>0.166</td>
<td>0.130</td>
<td>27.02</td>
<td>26.61</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>21</td>
<td>0.581</td>
<td>0.236</td>
<td>374.79</td>
<td>196.07</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>8</td>
<td>0.065</td>
<td>0.070</td>
<td>16.94</td>
<td>17.02</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>10</td>
<td>0.155</td>
<td>0.135</td>
<td>40.20</td>
<td>49.54</td>
</tr>
</tbody>
</table>

### Table 5.3  Means for AR (7) Parameters for One Day (Φ₁) and Seven Days (Φ₇)

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>N</th>
<th>Presence of transaction</th>
<th>Transaction amount</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>Φ₁</td>
<td>Φ₇</td>
<td>Φ₁</td>
<td>Φ₇</td>
</tr>
<tr>
<td>Cluster 1</td>
<td>253</td>
<td>0.008714</td>
<td>0.010184</td>
<td>0.001245</td>
<td>0.007436</td>
</tr>
<tr>
<td>Cluster 2</td>
<td>50</td>
<td>-0.01257</td>
<td>-0.07441</td>
<td>-0.0018</td>
<td>-0.0799</td>
</tr>
<tr>
<td>Cluster 3</td>
<td>4</td>
<td>0.078857</td>
<td>0.187838</td>
<td>0.011265</td>
<td>0.028232</td>
</tr>
<tr>
<td>Cluster 4</td>
<td>311</td>
<td>0.006714</td>
<td>0.056576</td>
<td>0.000959</td>
<td>0.019623</td>
</tr>
<tr>
<td>Cluster 5</td>
<td>341</td>
<td>0.051286</td>
<td>0.045393</td>
<td>0.007327</td>
<td>0.028947</td>
</tr>
<tr>
<td>Cluster 6</td>
<td>21</td>
<td>0.025857</td>
<td>0.057036</td>
<td>0.003694</td>
<td>0.004974</td>
</tr>
<tr>
<td>Cluster 7</td>
<td>8</td>
<td>0.078286</td>
<td>0.147415</td>
<td>0.011184</td>
<td>-0.02234</td>
</tr>
<tr>
<td>Cluster 8</td>
<td>10</td>
<td>0.134714</td>
<td>0.110601</td>
<td>0.019245</td>
<td>0.07818</td>
</tr>
</tbody>
</table>
Clusters 3, 7, and 8 were regarded as outlier clusters due to their low frequencies. These outlier clusters however could be of great importance if we had more details about the customers. For example if we were provided with the payment patterns of the customers, we could have identified risky outlier clusters. Banks could monitor customers who enter these clusters and come up with strategies to minimize their loss. In the pilot study we do not deal with these outlier clusters because of lack of information. Interestingly the three outlier clusters all have relatively high values for the $\Phi_7$ on both series suggesting that people in these clusters tend to shop on a weekly basis.

The other five clusters were given the following names on the basis of the summary data shown in Tables 5.2.

Cluster 1: Rare spenders
Cluster 2: Blue Moon spenders
Cluster 4: $2000 per month
Cluster 5: $800 per month
Cluster 6: Big spenders

Our aim in this analysis is to identify patterns based on transaction amounts and group customers based on similar behaviour. The above 5 clusters give us 5 different groups of customers with similar behaviour based on transactions over 6 months using AR parameters and moments extracted from the two transaction time series. Note that only the daily transaction amounts were needed to identify the characteristics of each cluster. Using linear discriminant analysis and ignoring the above outlier clusters, classification rules were developed for the above five clusters using parameters derived for each customer from their daily time series data.

Assumption of normality and equality of covariance matrices in the population were reasonably valid given the small sample size.
Study 1 - Part b) Monthly time series:

In the above analysis six months of transactions were time series each of 183 time points. In this analysis daily transaction amounts and presence of transactions were considered for each of the 1000 customers, with 6 separate dataset representing each of the six months. Each of these datasets includes 1000 customer’s daily transactions, daily presence of transaction, id and date. Daily transactions for each customer for a month were considered as time series. Also daily presence (or absence) of transaction for each customer for one month was considered as a second time series.

The analysis is the same as before. An AR (7) model is fitted to each series in order to extract 7 time series parameters. Mean, standard deviation, kurtosis and skewness are also calculated for each of the series for each month.

Using the classification rules developed in Part a, the behaviour of each customer within each month was used to classify each customer on a monthly basis. A 6 digit code is created for each customer to show their expenditure patterns over the 6 month period. The clusters names are the same as the clusters in the previous study which are as follows:

Cluster 1: Rare spenders
Cluster 2: Blue Moon spenders
Cluster 4: $2000 per month
Cluster 5: $800 per month
Cluster 6: Big spenders

Table 5.4 shows monthly cluster numbers and the resulting six digit code for a random sample of 21 customers.
Table 5.4  
Monthly Cluster Classifications for Randomly Selected 10 Customers

<table>
<thead>
<tr>
<th>Id</th>
<th>June</th>
<th>July</th>
<th>August</th>
<th>September</th>
<th>October</th>
<th>November</th>
<th>Cluster code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>444544</td>
</tr>
<tr>
<td>2</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>444444</td>
</tr>
<tr>
<td>3</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>444444</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>444444</td>
</tr>
<tr>
<td>5</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>4</td>
<td>455545</td>
</tr>
<tr>
<td>6</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>245444</td>
</tr>
<tr>
<td>7</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>544444</td>
</tr>
<tr>
<td>8</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>445555</td>
</tr>
<tr>
<td>9</td>
<td>5</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>5</td>
<td>5</td>
<td>554455</td>
</tr>
<tr>
<td>10</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>4</td>
<td>444444</td>
</tr>
</tbody>
</table>

For example, for id 1, the June expenditure pattern suggests cluster four, which on an average has a mean daily transaction amount of 68 dollars. The expenditure pattern remains similar for the months July and August. However, in September, the customer moves to cluster five, which has a mean daily transaction amount of 27 dollars. This shows a change in customer’s expenditure pattern. However, in the next two months, he moves back to cluster four.

Monitoring movements of customers between clusters is useful to the banks with respect to risk and marketing perspective. Identifying risky clusters will help a bank identify and deal with risky customers in a timely and appropriate fashion. Clustering is very useful with marketing perspective. Understanding different clusters properties, appropriate strategies can be made to increase the profit. For example in this analysis, cluster 6 represents big spenders. These customers are profitable for a bank. Banks can reward customers in this cluster encouraging them to spend more. On the other hand, banks can come up with strategies which are beneficial for customers in cluster 2 (rare spenders), to encourage them to use their credit card more often.

5.4.2 Study 2 (clustering based on merchant category code).

The previous study helps us understand the monthly behaviour pattern of customers and any change in their spending patterns. This could provide banks with a good system to understand the transaction patterns exhibited by good and bad customers. If banks understand the customers’ spending and payment behaviour, such as how much they spend
spend, where they spend, how regularly they spend and pay; banks could better determine the financial circumstances of their customers. This could help banks come up with appropriate strategies for different groups of customers, helping banks reduce loss and improve profitability.

To further understand a customer’s financial circumstances, study 2 was conducted on the same 1000 customers using merchant details. Here we were interested to determine the main type of spending for each customer. This is useful in identifying customers who are likely to have financial problems. For example if a customer is spending regularly on flights or jewelry, it could indicate that he is in a good financial situation. But if the same customer changes his spending pattern and uses his credit card only on food (which is a necessity), this could suggest financial crisis. Looking at his payment behaviour, we could predict future outcomes.

5.4.2.1 Preparation of data
The variables used for this analysis are *anon_id, merchant's name, merchant category code (MCC), transaction amount and date*. Using the merchant’s names and the category code provided by the bank, we created a new code called merchant code (MC) to categorize these merchants into fewer categories. For example, merchants such as Emirates airline (mcc=3026), Air New Zealand (mcc=3025), Qantas (mcc=3012) were all assigned to a new merchant code known as “airlines” with MC=1. The data provided to us originally included 331 merchant category codes. There were also 64153 transactions with missing merchant category codes.

For this study we reduced the number of merchant categories to 15. The codes and the name for each code are described below:

- Airline MC1;
- Cash Advance MC2;
- Lodgings MC3;
- Car rental MC4;
- Transport MC5;
- Travel agent MC6;
- Auto related MC7;
For each customer we calculated the amount spent on each of these 15 categories per day. Transaction amount spent per day over 6-month on each of these 15 groups were taken as univariate time series. Different analysis was performed on this data to verify the usefulness of the analysis based on the aim of the project. The next section gives details of the two analyses based on transactions spent per product.

**Study 2 - Part a) Clustering of Customers Based on MC**

As in the previous study, raw data was not directly used for clustering. Instead parameters were extracted from each of the series. Looking at the MC’s, it does not make sense to consider the AR parameter for all of them. It is very likely that people spend money on supermarket or retail items every week, but it is not common to spend weekly on airlines or lodging. Therefore, we chose to consider AR parameters only for retail, supermarket, oil and petrol. However, moments for daily expenditure were calculated for all MC codes. For each customer we had 81 input variables (21 AR parameters and 60 moments). Note that series for daily presence of transactions were not considered because it seemed from Study 1 that these series were not necessary. The pseudo T-squared criterion suggested 8 clusters in this case when an Agglomerative Hierarchical Clustering with Ward’s method was applied to this data. Below is the dendrogram for the 1000 customers.
Table 5.5  Frequency Table for Each Cluster

<table>
<thead>
<tr>
<th>Cluster</th>
<th>Frequency</th>
<th>Percent</th>
<th>Cumulative Frequency</th>
<th>Cumulative Percent</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>626</td>
<td>62.66</td>
<td>626</td>
<td>62.66</td>
</tr>
<tr>
<td>2</td>
<td>286</td>
<td>28.63</td>
<td>912</td>
<td>91.29</td>
</tr>
<tr>
<td>3</td>
<td>78</td>
<td>7.81</td>
<td>990</td>
<td>99.10</td>
</tr>
<tr>
<td>4</td>
<td>2</td>
<td>0.20</td>
<td>992</td>
<td>99.30</td>
</tr>
<tr>
<td>5</td>
<td>3</td>
<td>0.30</td>
<td>995</td>
<td>99.60</td>
</tr>
<tr>
<td>6</td>
<td>2</td>
<td>0.20</td>
<td>997</td>
<td>99.80</td>
</tr>
<tr>
<td>7</td>
<td>1</td>
<td>0.10</td>
<td>998</td>
<td>99.90</td>
</tr>
<tr>
<td>8</td>
<td>1</td>
<td>0.10</td>
<td>999</td>
<td>100.00</td>
</tr>
</tbody>
</table>

Frequency, percentage frequency and the cumulative frequency of customers for the 8 clusters are shown in Table 5.5. Clearly 63% of the customers are in cluster 1, and 91% are in cluster 1 and 2. Having 91% of customers in 2 clusters does not help us differentiate between customers. This may be because too many merchant codes have been considered or because the sample size is too small.
Study 2 - Part b) Clustering of Customers Based on Percentage Changes in Monthly Expenditure on just a few MC

However, it may be that we should rather be looking at changes in expenditure patterns. The change in expenditure pattern could provide an early indication of financial trouble for customers. To understand the changes in financial situation of customers, 5 new merchant category groups (MCGs) were created. The groups are selected in such a way that we can differentiate between a financially stable customer and a customer under financial stress. The groups are as follows:

- MCG1 Airlines, lodging, travel agent
- MCG8 Supermarket
- MCG10 Oil and petrol
- MCG13 Retail (households, electronics)
- MCG14 Luxury (shoes, cloths, beauty salon, gym, etc)

For example, a customer who spends a lot on airline, travel agents, gym and beauty salon would probably not be under financial stress. On the other hand, a customer who spends only on food and petrol, which are the necessities in life, could be under financial stress. Having information on the payment behaviour of these customers, their financial circumstances could be identified. For this study we dropped AR parameters because we have too little data for each Merchant Code. Instead only monthly expenditure for each MCG code was considered.

The total amount spent for each MCG was calculated for each month, for each customer. From that, the percentage change for each MCG from one month to the next was determined for each customer from June to November. Therefore our new data for each month included 5 variables for each customer which gave the percentage change on each MCG for one month to another. Table 5.6 shows the monthly percentage change in MCG expenditure for 10 customers.
Table 5.6  
*Percentage Changes in Amount Spent for Each MCG for a Single Month*

<table>
<thead>
<tr>
<th>Customer id</th>
<th>Travel</th>
<th>Supermarket</th>
<th>Oil/petrol</th>
<th>Retail</th>
<th>Luxury</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>-15.31%</td>
<td>0.00%</td>
<td>-16.63%</td>
<td>0.00%</td>
<td>-15.31%</td>
</tr>
<tr>
<td>2</td>
<td>27.89%</td>
<td>-1.54%</td>
<td>-7.67%</td>
<td>-10.40%</td>
<td>27.89%</td>
</tr>
<tr>
<td>3</td>
<td>4.72%</td>
<td>0.94%</td>
<td>16.55%</td>
<td>-4.30%</td>
<td>4.72%</td>
</tr>
<tr>
<td>4</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>5</td>
<td>14.20%</td>
<td>1.56%</td>
<td>15.18%</td>
<td>-32.85%</td>
<td>14.20%</td>
</tr>
<tr>
<td>6</td>
<td>-2.69%</td>
<td>5.17%</td>
<td>40.11%</td>
<td>-11.77%</td>
<td>-2.69%</td>
</tr>
<tr>
<td>7</td>
<td>-5.43%</td>
<td>0.00%</td>
<td>-26.15%</td>
<td>29.11%</td>
<td>-5.43%</td>
</tr>
<tr>
<td>8</td>
<td>0.00%</td>
<td>31.68%</td>
<td>-9.81%</td>
<td>-22.86%</td>
<td>0.00%</td>
</tr>
<tr>
<td>9</td>
<td>-6.86%</td>
<td>12.34%</td>
<td>-4.13%</td>
<td>-3.38%</td>
<td>-6.86%</td>
</tr>
<tr>
<td>10</td>
<td>7.71%</td>
<td>10.96%</td>
<td>3.30%</td>
<td>4.57%</td>
<td>7.71%</td>
</tr>
</tbody>
</table>

Based on the 5 variables given in the Table above (excluding customer id and date), customers were classified into clusters, using Agglomerative Hierarchical Clustering with Ward’s method. According to the pseudo $T$-squared statistic the optimum number of clusters is 12. The most common clusters are cluster 1 and 11 and the dendrogram based on percentage change in monthly expenditure is shown in Figure 5.4.

![Dendrogram based on percentage change in monthly expenditure for five MC codes.](image)

A special five letter code is assigned to each cluster according to the average percentage increase or decrease in the amount spent for each MCG in each month. The first letter in
this code refers to Airlines, lodging, travel agent (MCG1), the second letter refers to supermarket expenditure (MCG8), the third letter refers to oil and petrol expenditure (MCG10), the fourth letter refers to household retail (MCG13) and the last letter refers to other more personal expenditure (MCG14). Capital letters indicate an increase in expenditure and lower case letters indicate a decrease in expenditure in any month.

<table>
<thead>
<tr>
<th>Code</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>N/n</td>
<td>(0 - 9.99)%</td>
</tr>
<tr>
<td>A/a</td>
<td>(10- 19.99)%</td>
</tr>
<tr>
<td>B/b</td>
<td>(20- 50)%</td>
</tr>
<tr>
<td>D/d</td>
<td>(&gt; 50)%</td>
</tr>
</tbody>
</table>

For example, cluster 2 has a code (nnnDN) which indicates, there was a (0-9.99) % decrease on MCG1, MCG8 and MCG10. In addition, there was a >50% increase on MCG13 and a (0-9.99%) increase on MCG14.

Table 5.7 shows the average monthly percentage change for each MCG group for each cluster over the period of the first month (June to July). Cluster 1 is the largest cluster with almost 40% of all customers. In this cluster expenditure on each of the MCGs decrease or increase, on average by 0-10%. This change is very small indicating customers’ expenditure in July is similar to their June expenditure.
Table 5.7  *Average Monthly Percentage Change in Expenditure by MC Group from June to July and Frequencies for Each Cluster*

<table>
<thead>
<tr>
<th>Cluster no.</th>
<th>Freq</th>
<th>Travel</th>
<th>Supermarket</th>
<th>Oil/petrol</th>
<th>Retail</th>
<th>Luxury</th>
<th>code</th>
<th>cluster name</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1955</td>
<td>-2.6%</td>
<td>-2.0%</td>
<td>-0.8%</td>
<td>1.5%</td>
<td>0.0%</td>
<td>n n N</td>
<td>A</td>
</tr>
<tr>
<td>2</td>
<td>366</td>
<td>-5.0%</td>
<td>-5.4%</td>
<td>-8.0%</td>
<td>59.2%</td>
<td>0.2%</td>
<td>n n D N</td>
<td>B</td>
</tr>
<tr>
<td>3</td>
<td>292</td>
<td>-84.1%</td>
<td>2.7%</td>
<td>1.5%</td>
<td>19.7%</td>
<td>4.1%</td>
<td>d N N A</td>
<td>C</td>
</tr>
<tr>
<td>4</td>
<td>455</td>
<td>71.4%</td>
<td>-5.6%</td>
<td>-0.8%</td>
<td>-12.3%</td>
<td>-9.8%</td>
<td>D n n a</td>
<td>D</td>
</tr>
<tr>
<td>5</td>
<td>92</td>
<td>-0.1%</td>
<td>8.4%</td>
<td>8.4%</td>
<td>0.8%</td>
<td>-93.0%</td>
<td>n N N d</td>
<td>E</td>
</tr>
<tr>
<td>6</td>
<td>316</td>
<td>-16.4%</td>
<td>-6.0%</td>
<td>-4.6%</td>
<td>-1.9%</td>
<td>55.4%</td>
<td>a n D a</td>
<td>F</td>
</tr>
<tr>
<td>7</td>
<td>134</td>
<td>2.0%</td>
<td>0.5%</td>
<td>-0.6%</td>
<td>-87.1%</td>
<td>-1.3%</td>
<td>N n d a</td>
<td>G</td>
</tr>
<tr>
<td>8</td>
<td>79</td>
<td>-13.6%</td>
<td>65.7%</td>
<td>-1.6%</td>
<td>-19.6%</td>
<td>1.3%</td>
<td>a D n a</td>
<td>H</td>
</tr>
<tr>
<td>9</td>
<td>81</td>
<td>-4.1%</td>
<td>0.1%</td>
<td>-1.5%</td>
<td>78.7%</td>
<td>-68.8%</td>
<td>n N D d</td>
<td>I</td>
</tr>
<tr>
<td>10</td>
<td>215</td>
<td>0.6%</td>
<td>0.4%</td>
<td>1.0%</td>
<td>-53.9%</td>
<td>55.8%</td>
<td>N N D D</td>
<td>J</td>
</tr>
<tr>
<td>11</td>
<td>577</td>
<td>4.5%</td>
<td>6.1%</td>
<td>6.6%</td>
<td>-20.0%</td>
<td>2.1%</td>
<td>N N N a</td>
<td>K</td>
</tr>
<tr>
<td>12</td>
<td>438</td>
<td>2.6%</td>
<td>1.5%</td>
<td>1.2%</td>
<td>15.4%</td>
<td>-30.1%</td>
<td>N N A b</td>
<td>L</td>
</tr>
</tbody>
</table>

The clusters were given names based on the average percentage changes in the amount of spent on each MCG. These names are as follows:

- **Cluster A**  Normal  (increase or decrease of 0-10%)
- **Cluster B**  Better house  (MCG13 ↑)
- **Cluster C**  Better life  (MCG8, MCG10, MCG13, MCG14 ↑)
- **Cluster D**  Holiday  (MCG14 ↑)
- **Cluster E**  Self-sacrifice  (MCG14 , MC1 ↓)
- **Cluster F**  Self centered  (MCG14 ↑, all others ↓)
- **Cluster G**  Bon vivant  (MCG1, MCG8 ↑, all others ↓)
- **Cluster H**  Family expansion  (MCG8 ↑)
- **Cluster I**  Less social  (MCG1, MCG10, MCG14 ↓)
- **Cluster J**  Preset life  (MCG 13 ↓, MCG 14 ↑)
- **Cluster K**  Social  (MCG13 ↓)
- **Cluster L**  Family guy  (MCG14 ↓)
For example the name for cluster L was chosen to reflect the reduction in personal luxury expenditure in this month. Considering the above twelve clusters; we developed a rule for classifying customers into these 12 clusters based on their June-July transaction behaviour. Linear discriminant analysis was used to develop classification rules which allowed the expenditure behaviour of each customer within each month to be assigned to one of the above clusters. Assumption of normality was reasonable for this data. Assumption of equal covariance matrices in the population was also reasonably valid given the small sample size.

Therefore based on a monthly percentage change in expenditure for each of the above five MCG’s, customers were assigned a cluster for each of the five months resulting in another 5 letter code, showing the change in expenditure patterns for each customer over the 6 month period. Table 5.8 shows monthly cluster names and the resulting five digit code for the first 15 customers.

<table>
<thead>
<tr>
<th>Id</th>
<th>jun/jul</th>
<th>jul/ aug</th>
<th>aug/ sep</th>
<th>sep/oct</th>
<th>oct/nov</th>
<th>Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>B</td>
<td>G</td>
<td>D</td>
<td>A</td>
<td>A B G D A</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>K</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>I K A A A</td>
</tr>
<tr>
<td>3</td>
<td>C</td>
<td>B</td>
<td>K</td>
<td>K</td>
<td>A</td>
<td>C B K K A</td>
</tr>
<tr>
<td>4</td>
<td>K</td>
<td>B</td>
<td>J</td>
<td>A</td>
<td>A</td>
<td>K B J A A</td>
</tr>
<tr>
<td>5</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>B</td>
<td>G</td>
<td>C A A B G</td>
</tr>
<tr>
<td>6</td>
<td>A</td>
<td>A</td>
<td>K</td>
<td>B</td>
<td>K</td>
<td>A A K B K</td>
</tr>
<tr>
<td>7</td>
<td>K</td>
<td>A</td>
<td>D</td>
<td>B</td>
<td>K</td>
<td>K A D B K</td>
</tr>
<tr>
<td>8</td>
<td>K</td>
<td>F</td>
<td>E</td>
<td>B</td>
<td>K</td>
<td>K F E B K</td>
</tr>
<tr>
<td>9</td>
<td>A</td>
<td>D</td>
<td>C</td>
<td>A</td>
<td>A</td>
<td>A D C A A</td>
</tr>
<tr>
<td>10</td>
<td>A</td>
<td>K</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A K A A A</td>
</tr>
<tr>
<td>11</td>
<td>C</td>
<td>D</td>
<td>A</td>
<td>D</td>
<td>C</td>
<td>C D A D C</td>
</tr>
<tr>
<td>12</td>
<td>D</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>H</td>
<td>D A A A H</td>
</tr>
<tr>
<td>13</td>
<td>A</td>
<td>F</td>
<td>E</td>
<td>A</td>
<td>A</td>
<td>A F E A A</td>
</tr>
<tr>
<td>14</td>
<td>A</td>
<td>K</td>
<td>A</td>
<td>K</td>
<td>A</td>
<td>A K A K A</td>
</tr>
<tr>
<td>15</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A</td>
<td>A A A A A</td>
</tr>
</tbody>
</table>

Consider a customer with a five cluster code of (JEAAA). This customer is initially in cluster J (N N N d D). This indicates that in July, they increased their expenditure by 0-
10% on MC1, MC8 and MC10. They spend 50% more on MC 15 and 50% less on MC14 compared to June. In August they moved into cluster E (n N N N d) , reducing their expenditure on MCG1 by 0-10% and on MCG15 by more than 50%. Expenditure on MCG8, MCG10, MCG13 increased by 0-10%. In the next 3 months they stayed in cluster A (nnNN). So, over this 6 month period they moved from the “social” cluster (a fling) to the “self-sacrifice” (redemption) cluster before moving to the “normal” cluster for the next 3 months.

5.5 Discussion

In this study transaction behaviour of customers are monitored by classifying customers into meaningful clusters containing customers with similar spending behaviour. Time series clustering is performed to cluster customers initially based on transaction amounts and then based on five types of expenditure (example supermarket, petrol, airline etc) and then monthly percentage change in expenditure in these five categories. Clustering is a useful tool for marketing. Appropriate strategies can be made based on the characteristics of clusters to increase a banks’ profit. For example, banks could come up with loyalty programs for customers in good clusters; rewarding customers for the regular use of their credit cards.

Movements of customers between clusters are monitored over 6 month period to identify changes in the financial situation of customers. A methodology is developed for monitoring these changes in customer behaviour to provide early warning of deterioration in a customer’s financial circumstances. A code is created to explain the movements between clusters and to get an understanding of the changes in each customer’s financial situation. Understanding changes in financial situation of customers is advantageous for banks in making better decisions in managing existing customers. For example consider a customer with movement codes HIJKK and MAAMA in two successive 6 month periods. This indicates that, in the first 6 months, the percentage change of the customers’ expenditure on average increased by 50% or more for some MCs, while for other MCs it increased by 0-10% on average. However, in the next 6 months, he moves into clusters M and A suggesting that he is mainly
spending on supermarket and petrol, which are the necessities. This could be due to deterioration in the customer’s financial situation. If we incorporate payment behaviour into the decision process, we should get even more interesting conclusions.

The aim of this pilot study was to gain more understanding of customers’ transaction behaviour. The dataset used in this study is limited to transaction amounts and Merchant category codes. No information about payment amounts or current status of customers (risky/non-risky) was provided. Despite the limited information available, meaningful clusters have been created and the movements of customers between clusters can be monitored.

Clustering is not directly applied to the time series data. Instead the clustering is performed on the model parameters and moments extracted from the time series to group together the customers with similar expenditure behaviour. The usefulness of these time series clusters cannot be evaluated in terms of delinquency prediction because we don’t have information about the risk status of the customers in this study. This will be done in the main study.

This pilot study has laid the foundation for the analyses described in the next chapter. Despite a very small data set (1000 customers) it has been possible to demonstrate that clustering of daily transactional time series for a single month can be clustered in a meaningful way. The main study involves a much larger data set allowing a much more rigorous analysis.
Chapter 6

MAIN STUDY- DATA PREPARATION

Truth is ever to be found in the simplicity, and not in the multiplicity & confusion of things.

- English physicist, mathematician, astronomer, alchemist Sir Isaac Newton
Table 6.1  
*Outline of chapter six*

<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>6.1 Introduction</td>
</tr>
<tr>
<td>6.2 Preparation of data for main study</td>
</tr>
<tr>
<td>6.2.1 Transaction data</td>
</tr>
<tr>
<td>6.2.1.1 Transaction amount</td>
</tr>
<tr>
<td>6.2.1.2 Merchant category code</td>
</tr>
<tr>
<td>6.2.2 Payment data</td>
</tr>
<tr>
<td>6.2.3 Delinquency data</td>
</tr>
<tr>
<td>6.2.4 Input variables</td>
</tr>
</tbody>
</table>

**6.1 Introduction**

Credit card customers default mainly due to changes in their financial circumstances and changes in the financial circumstances of customers’ results in unusual transaction behaviour. We believe that if a customer is financially sound, his transaction behaviour would be different compared to a customer who is financially stressed. To capture these changes in behaviour, customers’ transactions can be monitored based on what products they buy, how often they use their card and how the credit is re-paid. Based on type of products, amounts spent per product, how often credit card is used and amount repaid, customers are grouped into clusters and analysed to predict the probability of risky customers.

In the previous chapter concerning the pilot study, the dataset was limited only to transaction amount and merchant category codes (MCCs). Delinquency (overdue) status and repayment amounts were not provided; therefore we were unable to come up with a model for delinquency. In the main study, additional to transaction amounts and MCCs, we were provided with delinquency status and repayment amounts. With this additional information on delinquency status and payment amounts we are able to develop a scoring model for delinquency and we were able to validate this model because there is plenty of data.
It has been found in the literature that the use of new and more predictive variables can improve the performance of scorecards (Hand & Henley, 1997). In this research we have created new variables which we believe can differentiate between financially well and financially stressed customers. Creation of these new predictive variables is a primary contribution of this research.

This chapter explains the steps in creating the new variables from the available data. The data for the main study comes from a leading bank of Australia. This bank provided us with datasets which allow us to create many variables which proved to be very useful in the prediction of delinquency. The methods developed in the pilot study are used in this study. As was the case in the pilot study the daily transaction time series are not directly used. Autoregressive (AR) models are fitted to each daily transaction time series in order to extract AR parameters. Along with the AR parameters, moments are extracted from these time series. Also product based transaction amounts are considered as input variables. In the previous chapter only a few Merchant Categories Groups (MCG) were considered since the dataset was very small. The dataset used in this study is much larger. There were originally 902 Merchant Category Codes in this study which are grouped based on the similarities of services or products. These groups are then used to define MCG specific transactional data as input variables. The outline of the chapter is shown in Table 6.1.

6.2 Preparation of data for the main study:

Banking data is multidimensional and very large in size. A leading bank of Australia has provided 60 GB of transactional and payment data spanning the period October 2006-September 2008 comprised of approximately 2.4 million unique accounts for this study. In addition we were provided with delinquency data containing the delinquency status of approximately 1.5 million customers over the same period. Three different datasets namely transactional data, payment data and delinquency data were provided to us. These datasets are explained in the following section.
6.2.1 Transactional data:

As indicated in table 6.2, the transactional data includes 5 variables. The name and type of the variables are given in the Table below followed by their descriptions.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acct_Num</td>
<td>Num</td>
</tr>
<tr>
<td>Date_Of_Transaction</td>
<td>Char</td>
</tr>
<tr>
<td>Merchant_Category_Code</td>
<td>Char</td>
</tr>
<tr>
<td>Merchant_Terminal_Name</td>
<td>Char</td>
</tr>
<tr>
<td>Transaction_code</td>
<td>Num</td>
</tr>
<tr>
<td>Transaction_Amount</td>
<td>Num</td>
</tr>
</tbody>
</table>

**Acct_num** : This is the customer id provided by the bank. This id is different from the actual account number. The bank has created an account link which links the acct_num to the actually account. We were not provided with the account number or the account link. This meant that the data was de-identified in order to preserve confidentiality.

**Date_of_transaction** : The date when the transaction took place. Valid date in CCYYMMDD format (cannot be zero).

Where:

- CC = Century
- YY = Year
- MM = Month
- DD = Day

For example, the 31st of May 2001 is '20010531'.
**Merchant_category_code:** Each merchant (or retailer) is classified by a code which determines their type of business. The merchant is assessed by the acquiring bank\(^9\) when it takes on the credit card facility. This field identifies the merchant's category code. It is a 4 digit code when populated, 3 digit codes have a leading zero.

**Merchant_terminal_name:** The name of the merchant.

**Transaction_code:** Identifies the type of transaction being processed. Transaction codes are often used in conjunction with Transaction Reason Codes. Transaction Codes are 3 digits. Some examples are as follows:

- 005 – Purchase
- 007 – Cash advance
- 065 – Payment
- 069 – Non-monetary (to be ignored for most purposes)

**Transaction_amount:** The transaction amount in local currency. This amount is posted to the cardholder account.

The variables in Table 6.2 are not directly used for the analysis. Based on the transactional data, several variables were extracted for our analysis. The procedure for extraction and the variable names are given in the following section.

### 6.2.2 Extracted variables from Transaction amount:

The transaction amount included payments, transactions and fees/fines. To start with, transactions were separated from the payments and fees/fines were excluded. The customer transactional amounts were aggregated to a daily level. If there were no transaction on particular day, then the transaction amount for this day was set at zero,

---

\(^9\) An **acquiring bank** is the bank that accepts credit (or debit) card payments for products or services on behalf of a merchant.
otherwise it was set to the amount spent on the credit card on that day. If there were more than one transaction per day, transaction amount was the sum of all the transactions on that day. For each account these daily transaction amounts were used to create a separate time-series for each month.

As in the previous chapter, time series were not directly used in this analysis. The time series patterns were observed in these series. In most cases a weekly pattern (lag of 7) was noted. Autoregressive models with lag 7 {AR (7)} were fitted to each of the series, allowing the estimation of the autoregressive parameters (Ø₁, Ø₂, ..., Ø₇). Moments (mean, standard deviation, kurtosis, and skewness) were also calculated for each time series. Therefore 11 variables (refer Table 6.3) were extracted from each transaction time series for each month.

Table 6.3  Extracted Variable Names and Descriptions Based on Transaction Amount

<table>
<thead>
<tr>
<th>Variable names</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>AR_1</td>
<td>Autoregressive parameter (lag 1)</td>
</tr>
<tr>
<td>AR_2</td>
<td>Autoregressive parameter (lag 2)</td>
</tr>
<tr>
<td>AR_3</td>
<td>Autoregressive parameter (lag 3)</td>
</tr>
<tr>
<td>AR_4</td>
<td>Autoregressive parameter (lag 4)</td>
</tr>
<tr>
<td>AR_5</td>
<td>Autoregressive parameter (lag 5)</td>
</tr>
<tr>
<td>AR_6</td>
<td>Autoregressive parameter (lag 6)</td>
</tr>
<tr>
<td>AR_7</td>
<td>Autoregressive parameter (lag 7)</td>
</tr>
<tr>
<td>Mean</td>
<td>Average amount of transactions spend per month</td>
</tr>
<tr>
<td>Skewness</td>
<td>Skewness of the amount of transactions over the month</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>Kurtosis for the amount of transactions over the month</td>
</tr>
<tr>
<td>Standard deviation</td>
<td>Standard deviation of amount of transactions spend per month</td>
</tr>
</tbody>
</table>

6.2.3. Extracted variables from Merchant category code (MCC):

The raw dataset provided to us included 902 Merchant Category Codes (MCCs). These MCCs were grouped according to the product category. For example all food related
merchants were grouped into a category known as ‘C_food’; airline, hotel and holiday related merchants were grouped to form ‘C_travel’. Each of these 9 categories were then further investigated and subdivided into smaller groups to separate essential products from luxurious products under each category. For example under food related category ‘C_food’ there were many supermarket related merchants, different restaurants and bars. Therefore ‘C_food’ was subdivided into 3 different descriptions namely ‘supermarket’, ‘food’ and ‘eating/drinking out’. For example MCC 921 representing ‘Coles’, 922 representing ‘Safeway’ and 926 representing ‘Aldi’, were grouped into the same description group known as “Supermarket”. Different MCCs representing restaurant were categorized as “Food”. Bars and alcohol shops were categorized into “Eating/Drinking habit”. These description groups on their own are useful in identifying financially stressed customers. Therefore the 9 categories were subdivided into 26 description codes. The prefix C was used to differentiate between description groups and categories.

Creating the description groups and different categories allows a very useful way of differentiating between a financially stable customer and a financially stressed customer. For example a customer spending 270 dollar on his/ her credit card for dinner in a restaurant every week should not be identified as being similar to a customer who spends the same amount on alcohol and cigarettes every week. In the above example, if the customer spending on alcohol and cigarettes doesn’t repay for months, and a financially stable individual pays on time, our model might pick “Alcohol spending” as a significant variable of delinquency. Based on merchant category codes, 35 variables (description groups and categories) were created for this analysis. These variables are shown in the Table 6.4.
However, the categories were highly correlated to the description codes with the descriptions all contained within one of the categories. Therefore for the analysis only the descriptions have been used. Table 6.5 shows the total amount spent for the various description groups in March 2007. In this Table N is the number of MCCs which are combined to form each description group. The highest amount is spent on Retail which includes 23 different merchants.
Table 6.5  \textit{Total Amount Spent on Each Description During March 2007}

<table>
<thead>
<tr>
<th>Description</th>
<th>Total Amount</th>
<th>N</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial_Services</td>
<td>16206898319</td>
<td>11</td>
</tr>
<tr>
<td>Cash_Advance</td>
<td>10602426306</td>
<td>2</td>
</tr>
<tr>
<td>Small/Ambiguous</td>
<td>8169.8</td>
<td>4</td>
</tr>
<tr>
<td>Not_Retail</td>
<td>142495738.4</td>
<td>5</td>
</tr>
<tr>
<td>Car Rental</td>
<td>196707130.1</td>
<td>36</td>
</tr>
<tr>
<td>Public_Transport</td>
<td>214563903.7</td>
<td>6</td>
</tr>
<tr>
<td>Donations</td>
<td>350748165.9</td>
<td>5</td>
</tr>
<tr>
<td>Food</td>
<td>393380541</td>
<td>5</td>
</tr>
<tr>
<td>Health/Beauty</td>
<td>395697650.2</td>
<td>5</td>
</tr>
<tr>
<td>Hobby</td>
<td>712693661</td>
<td>12</td>
</tr>
<tr>
<td>Education</td>
<td>737103490</td>
<td>7</td>
</tr>
<tr>
<td>Sport/Rec</td>
<td>898186629.7</td>
<td>20</td>
</tr>
<tr>
<td>Holiday/Travel</td>
<td>990359646.5</td>
<td>9</td>
</tr>
<tr>
<td>Airline</td>
<td>1315763786</td>
<td>142</td>
</tr>
<tr>
<td>Clothing</td>
<td>1315952148</td>
<td>20</td>
</tr>
<tr>
<td>Computer/commercial</td>
<td>1445306124</td>
<td>12</td>
</tr>
<tr>
<td>Unclassified</td>
<td>1494961294</td>
<td>35</td>
</tr>
<tr>
<td>Drinking_Habit</td>
<td>1541551697</td>
<td>6</td>
</tr>
<tr>
<td>Lodging</td>
<td>1674657978</td>
<td>269</td>
</tr>
<tr>
<td>Household_Appliances</td>
<td>1989466906</td>
<td>19</td>
</tr>
<tr>
<td>Medical</td>
<td>2070062154</td>
<td>21</td>
</tr>
<tr>
<td>Construction/Maintenance</td>
<td>2145556654</td>
<td>25</td>
</tr>
<tr>
<td>Supermarket</td>
<td>2355451186</td>
<td>2</td>
</tr>
<tr>
<td>Utilities</td>
<td>2780540492</td>
<td>10</td>
</tr>
<tr>
<td>Automotive</td>
<td>2897187757</td>
<td>23</td>
</tr>
<tr>
<td>Retail</td>
<td>3196161344</td>
<td>23</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>58,063,888,871</strong></td>
<td><strong>734</strong></td>
</tr>
</tbody>
</table>
The percentage of the amount spent on each MCC per month was calculated for each individual customer. Then for each customer the percentage spent on each description group and category is calculated per month. To illustrate this more clearly an example is considered. Suppose an individual uses his/her credit card to buy a new motorcycle for $600 on 5\textsuperscript{th} of March. On 18\textsuperscript{th} March this individual spends $200 on a pair of shoes; $40 on petrol, and $40 on food; also every fortnight this individual pays $60 for gym facilities. The March transactional data provided to us for this individual is given in Table 6.6.

**Table 6.6**  \textit{March Transaction for Customer xxx Provided to us by the bank}

<table>
<thead>
<tr>
<th>Customer Id</th>
<th>Date</th>
<th>Amount</th>
<th>MCC</th>
</tr>
</thead>
<tbody>
<tr>
<td>xxx</td>
<td>05/03/2007</td>
<td>600.00</td>
<td>5571</td>
</tr>
<tr>
<td>xxx</td>
<td>15/03/2007</td>
<td>60.00</td>
<td>7997</td>
</tr>
<tr>
<td>xxx</td>
<td>18/03/2007</td>
<td>200.00</td>
<td>5661</td>
</tr>
<tr>
<td>xxx</td>
<td>18/03/2007</td>
<td>40.00</td>
<td>5983</td>
</tr>
<tr>
<td>xxx</td>
<td>18/03/2007</td>
<td>40.00</td>
<td>5814</td>
</tr>
<tr>
<td>xxx</td>
<td>30/03/2007</td>
<td>60.00</td>
<td>7997</td>
</tr>
</tbody>
</table>

The total expenditure of this customer during March 2007 is $1000. For our analysis the percentage of the amount spent for each description group by customer xxx in March 2007 is calculated as shown in Table 6.7.
Table 6.7  
*Percentages of Amounts Spent on Each Description Group*  

<table>
<thead>
<tr>
<th>Description</th>
<th>%</th>
<th>Description</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Financial_Services</td>
<td>0%</td>
<td>Airline</td>
<td>0%</td>
</tr>
<tr>
<td>Cash_Advance</td>
<td>0%</td>
<td>Clothing</td>
<td>20%</td>
</tr>
<tr>
<td>Small/Ambiguous</td>
<td>0%</td>
<td>Computer/Commercial</td>
<td>0%</td>
</tr>
<tr>
<td>Not_Retail</td>
<td>0%</td>
<td>Unclassified</td>
<td>0%</td>
</tr>
<tr>
<td>Car Rental</td>
<td>0%</td>
<td>Drinking_habit</td>
<td>0%</td>
</tr>
<tr>
<td>Public_Transport</td>
<td>0%</td>
<td>Lodging</td>
<td>0%</td>
</tr>
<tr>
<td>Donations</td>
<td>0%</td>
<td>Household_Appliances</td>
<td>0%</td>
</tr>
<tr>
<td>Food</td>
<td>4%</td>
<td>Medical</td>
<td>0%</td>
</tr>
<tr>
<td>Health/Beauty</td>
<td>0%</td>
<td>Construction/Maintenance</td>
<td>0%</td>
</tr>
<tr>
<td>Hobby</td>
<td>0%</td>
<td>Supermarket</td>
<td>0%</td>
</tr>
<tr>
<td>Education</td>
<td>0%</td>
<td>Utilities</td>
<td>0%</td>
</tr>
<tr>
<td>Sport/Rec</td>
<td>12%</td>
<td>Automotive</td>
<td>64%</td>
</tr>
<tr>
<td>Holiday/Travel</td>
<td>0%</td>
<td>Retail</td>
<td>0%</td>
</tr>
</tbody>
</table>

6.2.4 *Payment data:*

The payment dataset provided by the bank includes 6 variables. The name and type of variables are given in Table below.

Table 6.8  
*Payment Data*  

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acct_Num</td>
<td>Num</td>
</tr>
<tr>
<td>Balbt</td>
<td>Num</td>
</tr>
<tr>
<td>Balcash</td>
<td>Num</td>
</tr>
<tr>
<td>Stmtbal</td>
<td>Num</td>
</tr>
<tr>
<td>Stmtdte</td>
<td>Num</td>
</tr>
</tbody>
</table>
Acct_num: This is the customer id provided by the bank. This id is the common variable in every dataset provided to us by the bank and is explained in section 6.2.1.

Balbt: The amount of balance transferred from another credit card to the credit card under study.

Balcash: The amounts for cash advances (i.e. Cash withdrawn using credit card).

stmtbal: The amount owed by the customer to the bank on the statement date.

stmdte: The date when the statement was issued to the customer.

Since the number of customers in the database is very large, for processing ease the bank has divided them into groups with different statement dates. This means different payment dates. Our analysis includes monthly time series for which we wanted to compare the same transaction time sequence for all the customers. If we were to allow for different statement dates, we would be forced to have different models for the different statement dates. Therefore we chose to go for monthly analyses. Ignoring differences in payment date, we chose all transactions from the 1st of the month to the end of the month, even though the customers would not have the same statement date for payment. A few ratios based on balance and transaction amounts were created for our analysis for each customer in each month (see Table 6.9).

<table>
<thead>
<tr>
<th>Extracted Variables Based on Payment Data</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Extracted Variables</strong></td>
</tr>
<tr>
<td>Pay_Spend_Ratio</td>
</tr>
<tr>
<td>Pay_Bal_Ratio</td>
</tr>
<tr>
<td>Bal_Change_Bal_Ratio</td>
</tr>
<tr>
<td>End_start_bal_ratio</td>
</tr>
<tr>
<td>Balance_owed</td>
</tr>
</tbody>
</table>

Pay_spend ratio is the ratio of the monthly payment to monthly expenditure. If a customer pays as much as he spends then the pay_spend_ratio is equal to 1. If a customer’s payment and expenditure amounts are similar then this ratio is close to 1. A high pay_spend ratio indicates that the customer is paying higher amounts compared to
the amounts he is spending. A low pay_spend_ratio indicates that the customer is spending a higher amount compared to his repayment amount.

**Pay_bal_ratio** is the ratio of monthly payment to the overall balance owed at the end of the month. If a customer pays the amount of money he owes in full then the pay_spend_ratio is 1. If the customer pays most of the balance amount he owes then this ratio is close to 1. If a customer does not make any payments then the pay_bal_ratio is 0. A low pay_bal_ratio indicates that the customer is paying a low proportion of the amount he owes. A pay_bal_ratio greater than 1 indicates that the customer is paying more than the amount he owes. This is unlikely to happen. However in this dataset we had a few customers with pay_bal_ratio greater than 1. These customers were removed before the analysis since we didn’t have enough reasons to explain that pattern of behaviour.

**Bal_change_bal_ratio** is the ratio of balance at the end of the month to monthly expenditure. Values close to zero indicate that the amount of expenditure of a customer is the same as the amount he owes. Smaller values indicate that the customer is spending more than he owes and very high ratios indicate that the customer is spending much less than the amount he owes.

**Bal_change_bal_ratio** is the ratio of balance owed at the start of the month to balance owed at end of the month. If the balance owed at the start of the month is equal to the amount of the money owed at the end of the month, the bal_change_bal_ratio is 1. This could be because the customer does not spend any money and does not pay any money. Values less than 1 indicate that the balance has increased at the end of the month. Values greater than 1 indicates that the balance owed at the end of the month is less than the amount of balance owed at the start of the month.

**Balance_owed**: this is the balance owed by the customer to the bank on the statement date. This variable is the same as stmtbal, the variable which was provided to us by the Bank in the payment dataset.
6.2.5 Input variables

It has been found in the literature that the use of new predictive variables can improve the performance of scorecards (Hand & Henley, 1997). Creation of new predictive input variables is the main focus of this research. The commonly used behaviour scoring models use demographic variables for customers and variables which describe the long term financial behaviour of customers, such as time on books (describing how long the customer has been using his credit card) and the number of times a customer has become delinquent in the past one year (or 6 months). These variables are useful in prediction.

From the available datasets 45 input variables were extracted from transaction and payment data. Table 6.10 shows the list of extracted input variables.

Table 6.10 List of Input Variables

<table>
<thead>
<tr>
<th>Ar_1</th>
<th>Donations</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ar_2</td>
<td>Drinking_habit</td>
</tr>
<tr>
<td>Ar_3</td>
<td>Construction/Maintenance</td>
</tr>
<tr>
<td>Ar_4</td>
<td>Donations</td>
</tr>
<tr>
<td>Ar_5</td>
<td>Drinking_habit</td>
</tr>
<tr>
<td>Ar_6</td>
<td>Education</td>
</tr>
<tr>
<td>Ar_7</td>
<td>Financial_Services</td>
</tr>
<tr>
<td>Mean</td>
<td>Food</td>
</tr>
<tr>
<td>Standard Deviation</td>
<td>Health/Beauty</td>
</tr>
<tr>
<td>Skewness</td>
<td>Hobby</td>
</tr>
<tr>
<td>Kurtosis</td>
<td>Holiday/Travel</td>
</tr>
<tr>
<td>Pay_Spend_Ratio</td>
<td>Household_Appliances</td>
</tr>
<tr>
<td>Pay_Bal_Ratio</td>
<td>Lodging</td>
</tr>
<tr>
<td>Bal_Change_Bal_Ratio</td>
<td>Medical</td>
</tr>
<tr>
<td>End_start_bal_ratio</td>
<td>Not_Retail</td>
</tr>
<tr>
<td>Balance_Owed</td>
<td>Public_Transport</td>
</tr>
<tr>
<td>Airline</td>
<td>Retail</td>
</tr>
<tr>
<td>Automotive</td>
<td>Small/Ambiguous</td>
</tr>
<tr>
<td>Car Rental</td>
<td>Sport/Rec</td>
</tr>
<tr>
<td>Cash_Advance</td>
<td>Supermarket</td>
</tr>
<tr>
<td>Clothing</td>
<td>Unclassified</td>
</tr>
<tr>
<td>Computer/Commercial</td>
<td>Utilities</td>
</tr>
<tr>
<td>Construction/Maintenance</td>
<td></td>
</tr>
</tbody>
</table>
We were also provided with a delinquency dataset. This dataset was used to create the target variable for each study.

### 6.2.6 Delinquency data

As shown in Table 6.11 the delinquency data includes 3 variables.

<table>
<thead>
<tr>
<th>Variable Name</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Acct_Num</td>
<td>Num</td>
</tr>
<tr>
<td>Delq_Status</td>
<td>Char</td>
</tr>
<tr>
<td>Stmtdte</td>
<td>Num</td>
</tr>
</tbody>
</table>

*Acct_num:* is the customer id which is common in all the datasets.

*Delq_status:* indicates delinquency status of each customer. Delinquency status can take values of 0 to 5 indicating the number of missed payment. Missing more than 5 payments poses a serious threat to the bank. If a customer’s delinquency status is 1 it indicates that the minimum payment was missed for a month or a customer is one month overdue. If a customer has missed the minimum payment for more than 5 months, the delinquency status is recorded as “L”. The delinquency status for an inactive customer (for example, only one transaction every few months) is recorded as “E”. In summary delinquency status provided to us in the delinquency dataset is as follows:

- 0  no overdue payment
- 1  one month overdue
- 2  two months overdue
- 3  three months overdue
- 4  four months overdue
- 5  five months overdue
- L  more than five months overdue
- E  Inactive account
**Stmdtdte** is the statement date and the date on which delinquency status was last calculated.

Commonly customers are grouped based on three classes of risk (Hand & Henley, 1997) namely good risk, bad risk and indeterminate class. A good risk class includes customers who are not risky, bad risk class includes customers who are risky. Indeterminate class is a class where the risk is not yet known. In this study we consider a customer as a “good” customer if the customer is at most one month overdue (delinquency status 0 and 1). A “bad” customer is a customer who is more than 3 months overdue (delinquency status ≥3). An “indeterminate” customer is a customer who has missed payments for 2 month consecutively (delinquency status= 2).

Since credit scoring is all about classifying customers into good or bad risk, a common practice in the credit scoring industry is to exclude customers in the indeterminate class. Including the indeterminate class tends to reduce the accuracy of the model. Therefore all the indeterminate customers (customers who have missed two payments) are excluded from all datasets before analysis. Also inactive accounts (customers with delinquency status as “E”) were excluded too since these customers are irregular customers and including them in the model would cause inaccuracy.

### 6.3 Importance of the Variables

Before we start with the analysis, the strength of all the created variables were examined using a common method practiced by banks. Banks use Information value (IV) to examine the predictive power of the variables. To calculate the IV the continuous variables are binned into n groups (usually n=10 bins). The odds of a customer being good are calculated for each of these bins for each of the grouped attributes. For example a continuous age variable could be binned into 3 groups (< 25; 25-35; > 35). Then the odds of a customer being good in each of the three age group is calculated. The total strength of a variable or Information Value comes from information theory and is calculated using the formula:
\[ \sum_{i=1}^{n} \left( \frac{\text{Dist Good}_i - \text{Dist Bad}_i}{\text{Dist Bad}_i} \right) \times \ln \left( \frac{\text{Dist Good}_i}{\text{Dist Bad}_i} \right) \]

where \( \text{Dist Good} \) represents the proportion of good credit customers, \( \text{Dist Bad} \) represents the proportion of bad credit customers, \( \left( \frac{\text{Dist Good}}{\text{Dist Bad}} \right) \) measures the odds of being good, and \( (i=1,2,...,n) \) is the bin index of the grouped attribute being evaluated. Information value is always positive and variables with Information values of more than 0.3 are viewed as strong. Variables with Information Value of less than 0.1 are typically viewed as weak, however the weak variables could provide information when combined with other variables. Therefore the weak variables should not be discarded unless there is a valid reason. Mays (2004) recommends that even if the weak variables are not considered for the model, they should still be retained for further analysis.

Table 6.12 shows the calculated IV for the variables in this study.
Table 6.12  
*Information Value for Each Variable*

<table>
<thead>
<tr>
<th>Variable</th>
<th>IV</th>
<th>Variable</th>
<th>IV</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINANCIAL_SERVICES</td>
<td>1.907</td>
<td>EATING_DRINKING_OUT</td>
<td>0.47</td>
</tr>
<tr>
<td>bal_change_bal_ratio</td>
<td>1.7229</td>
<td>AR1_7</td>
<td>0.4528</td>
</tr>
<tr>
<td>end_start_bal_ratio</td>
<td>1.5697</td>
<td>CLOTHING</td>
<td>0.4307</td>
</tr>
<tr>
<td>C_NON_DISCRETIONARY</td>
<td>1.4674</td>
<td>CONSTRUCTION_MAINTENANCE</td>
<td>0.4086</td>
</tr>
<tr>
<td>pay_bal_ratio</td>
<td>1.1335</td>
<td>C_CONSTRUCTION_MAINTENANCE</td>
<td>0.4086</td>
</tr>
<tr>
<td>C_RETAIL</td>
<td>1.0305</td>
<td>C_TRAVEL</td>
<td>0.3183</td>
</tr>
<tr>
<td>sk</td>
<td>0.797</td>
<td>HOUSEHOLD_APPLIANCES</td>
<td>0.279</td>
</tr>
<tr>
<td>C_FOOD</td>
<td>0.7821</td>
<td>COMPUTER_COMMERCIAL</td>
<td>0.2595</td>
</tr>
<tr>
<td>RETAIL</td>
<td>0.7586</td>
<td>HEALTH_BEAUTY</td>
<td>0.2338</td>
</tr>
<tr>
<td>kr</td>
<td>0.7516</td>
<td>C_UNCLASSIFIED</td>
<td>0.2283</td>
</tr>
<tr>
<td>mean</td>
<td>0.7184</td>
<td>FOOD</td>
<td>0.2242</td>
</tr>
<tr>
<td>C_TRANSPORT</td>
<td>0.6795</td>
<td>SPORT_REC</td>
<td>0.219</td>
</tr>
<tr>
<td>AR1_5</td>
<td>0.6786</td>
<td>UNCLASSIFIED</td>
<td>0.2185</td>
</tr>
<tr>
<td>MEDICAL</td>
<td>0.6661</td>
<td>LODGING</td>
<td>0.2184</td>
</tr>
<tr>
<td>AUTOMOTIVE</td>
<td>0.6661</td>
<td>HOBBY</td>
<td>0.2002</td>
</tr>
<tr>
<td>pay_spend_ratio</td>
<td>0.6645</td>
<td>DONATIONS</td>
<td>0.1857</td>
</tr>
<tr>
<td>AR1_6</td>
<td>0.6271</td>
<td>EDUCATION</td>
<td>0.129</td>
</tr>
<tr>
<td>AR1_4</td>
<td>0.6151</td>
<td>AIRLINE</td>
<td>0.0738</td>
</tr>
<tr>
<td>AR1_2</td>
<td>0.6099</td>
<td>HOLIDAY_TRAVEL</td>
<td>0.0717</td>
</tr>
<tr>
<td>C_DISCRETIONARY</td>
<td>0.609</td>
<td>PUBLIC_TRANSPORT</td>
<td>0.0439</td>
</tr>
<tr>
<td>AR1_3</td>
<td>0.5853</td>
<td>CAR_RENTAL</td>
<td>0.0157</td>
</tr>
<tr>
<td>AR1_1</td>
<td>0.5746</td>
<td>NOT_RETAIL</td>
<td>0.0125</td>
</tr>
<tr>
<td>SUPERMARKET</td>
<td>0.5671</td>
<td>CASH_ADVANCE</td>
<td>0.0092</td>
</tr>
<tr>
<td>UTILITIES</td>
<td>0.5582</td>
<td>C_CASH_ADVANCE</td>
<td>0.0092</td>
</tr>
<tr>
<td>std</td>
<td>0.4813</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Except a few, all other variables have IV of greater than 0.1. Indeed many of the created variables have an IV greater than 0.3 representing a strong predictive power. Cash advance has the minimum IV. However, based on heuristic view, customers with high amount of cash advances are more likely to default. All the above variables are included for model building.
6.4 Discussion:

The data provided to us for the main study is limited to transaction amount, payment amount, merchant category codes and delinquency status. No other information such as postal code or other variables which are commonly used in building credit scoring model was provided.

From the available limited data, new variables are created in such a way that customers having financial difficulties could be recognized. These variables explain the current financial situation of the customers. We believe that a financially healthy customers’ spending pattern is different to a customer who is not financially sound; and that understanding the current financial behaviour of customers can help us in predicting the future behaviour of these customers.

The created new variables are used to build a behaviour scoring model in the next chapter. The performance of the model in predicting future behaviour of customers is used to verify the usefulness of the created new variables.
Chapter 7

MAIN STUDY - EXPLORATION OF CUSTOMERS’ BEHAVIOUR FOR FUTURE PREDICTION OF RISK

Old ways will always remain unless someone invents a new way and then lives and dies for it

- American writer, philosopher, artist Elbert Green Hubbard
### 7.1 Introduction

Monitoring customers’ transaction behaviour is a common practice in the banking industry. Using transaction behaviour, credit limit, account information and demographic information each customer is scored. Based on these scores banks can identify high risk and low risk customers. The general time line for monitoring transactional behaviour is 12-24 months (minimum time line is 6 months). During this time, many possible changes could have occurred in the customer’s financial behaviour. This is one of the disadvantages of current credit scoring systems (Thomas, 2000).

In this research, we have overcome this disadvantage by taking only one month’s transactional behaviour to predict the future risk. We believe that if a customer is financially sound, his transaction behaviour would be different compared to a customer who is financially under stress. To capture this behaviour, customers can be monitored based on what products they buy, how often they use their card and how the credit is repaid.

In the pilot study we had access to a small dataset with no payments data. Customers’ delinquency statuses were also not known which limited our analysis. The dataset used in this study is very large and it includes payment data and future delinquency status for the customers. Therefore we are able to build a behaviour scoring model predicting the future performance (good/bad) of the customers.
The outline of the chapter is shown in Table 7.1. In the next section we present the aim of the main study. Section 7.2 demonstrates the steps involved in preparing the input variables followed by data cleaning and sample generation in section 7.3. After cleaning the data, undersampling is performed to generate a sample including 20% future bad customers and 80% future good customers. This sample is partitioned into training, validation and test data. Two different models are applied to the training set and the performance of the models is assessed for forecasts over a 4-month time horizon. Section 7.4 explains the steps involved in creating the models and in model assessment. Using the test data the performance of the models is verified and the models are compared. Both the models are applied to the entire dataset including only 1.73% future bad customers to see if the models’ performance is reasonable when applied to a realistic but highly imbalanced dataset. Finally the model with a 6-month time horizon for forecasts is considered and compared with the models developed for 4-month time horizon forecasts.

### 7.2 Aim of the study

At this stage it is good to revise the definition of risky and non-risky customers. As mentioned in chapter one, we follow the Basel definition of default –minimum 90 days overdue. In this research a risky customer is referred to as a “future bad customer” indicating the customer has missed a minimum of 3 payments. A “future good customer” or a non-risky customer is a customer who has no overdue payments. The primary aim of this study is to monitor one month’s transactional behavior of customers and identify future risky customers in advance, so that the bank is aware of future risk and can come up with appropriate strategies before it is too late. A 4-month time horizon is initially used for prediction. In other words we aim to predict the probability of future bad customers **4 months in advance** using only one month’s transactional behaviour for customers. Two different models were applied to achieve this aim.

- **Model-1**: A conventional stepwise logistic regression.
- **Model-2**: A new hybrid SOM based classifier.
Logistic regression is by far the most common technique used for building credit scoring models according to many studies. A stepwise logistic regression is applied to select only the significant input variables for predicting future bad customers. In this study we propose a new hybrid SOM based classifier to predict the probability of future bad customers and the amount owed by these customers. The superiority of the proposed classifier is verified by comparing the performance of the two models. In the previous chapter the steps involved in variables creation were explained. In the following section the steps involved in building the behaviour scoring modes are explained.

7.3 Data cleaning and Sample generation

We aim to predict the probability of future bad customers 4 months in advance using only one months’ transaction behaviour of customers. We select one month because we want to understand the current financial situation of customers. External factors determined the month of data chosen to train our model. The emphasis was on selecting a month which is not an irregular month with respect to expenditure (not a festival month or holiday season). Year 2008 was a financially distressing year and hence it was avoided altogether. July is school holidays; November/December are festival months and January/February may be recovery time for holiday expenses. March 2007 was therefore chosen to predict the probability of future bad customers in July 2007.

7.4 A 4- Month Time Horizon

Based on March 2007’s transaction behaviour we aim to predict the “future performance” of the customers. Future performance is a binary target classifying customers as future good and future bad customers. A 4-months time horizon is selected for this study. So March 2007 data is used to predict the performance of customers in July 2007. In other words March 2007 is the observation period and the 4 months gap between April to July is the outcome period. If the outcome period is too long then the
observation period is too far in the past and therefore is not representative of future behaviour. However too short an outcome period will not allow sufficient time for an account to mature. That is, it takes a little time for the financial situation of a good customer to deteriorate to the extent that 3 payments are missed. Also too short an outcome period will not allow sufficient time for the bank to take preventative action.

Future performance is defined with “0” indicating a future good customer (i.e., a customer who is not overdue by July 2007) and with “1” indicating a future bad customer (i.e., a customer who has missed 3 or more payments by July 2007). Inactive customers and customers who missed payments for 1 or 2 months (indeterminate group) are excluded from all the studies (refer 6.2.6). Customers with missing delinquency status also are deleted. The transaction and payment behaviour of all remaining customers during March 2007 was monitored and the delinquency status of all these customers was checked for July 2007.

There were 1,003,515 customers who matched the above defined target including 17,406 future bad customers. This dataset contains 46 input variables including customer id (see Table 6.10) and a binary target variable to identify each customer as a future good or bad customer.

This dataset is imbalanced with a very low proportion of future bad customers. Modeling based on this imbalanced data with only 1.73 % future bad customers would classify all customers as future good customers and result in a high accuracy rate only for good customers. To get reliable results, a sample with approximately 20% future bad customers (target=1) and 80% future good customers was selected (refer Table 7.2). In other words under-sampling was performed selecting all the future bad customers and randomly selecting future good customers so that the proportion of future bad to future good customers is 1:4. Under-sampling resulted in a sample of size 86,030 with 20% future bad customers. Makuch (1999) makes the point that 100,000 good credit customers give sufficient information about the majority class so there is no need for more information on good customers.
Table 7.2  
*Frequency and Percentage of Good/Bad Customers*

<table>
<thead>
<tr>
<th>Frequency (%)</th>
<th>Future good customers</th>
<th>Future bad Customers</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entire Dataset</td>
<td>986,109 (98.27%)</td>
<td>17,406 (1.73%)</td>
<td>1,003,515</td>
</tr>
<tr>
<td>Random sample</td>
<td>69,624 (80%)</td>
<td>17,406 (20%)</td>
<td>86,030</td>
</tr>
</tbody>
</table>

7.4.1 Model development and validation

A random sample of data from March 2007 with 20% future bad customer is selected for analysis. The dataset is subdivided into training (40%), validation (30%) and test datasets (30%) before model development commences. Two models are applied to this data to predict the probability of future bad customers in July 2007 (see Figure 7.1). The performances of these models are compared and the superior model is selected as the final model.

![Figure 7.1: Procedures applied to the sample dataset](image-url)
Model-1: Stepwise Logistic Regression for classification

The most commonly used technique for building a credit scoring model is logistic regression. It is well-understood and easily interpretable. To check the usefulness and predictive capability of the created input variables, stepwise logistic regression was applied and the prediction accuracy of the model was estimated using the test data. The significant variables selected using stepwise logistic regression, are shown in Table 7.3.

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.56</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>HEALTH_BEAUTY</td>
<td>-0.06</td>
<td>0.0172</td>
</tr>
<tr>
<td>PAY_SPEND_RATIO</td>
<td>-0.025</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AR_5</td>
<td>-0.025</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>-0.024</td>
<td>0.0044</td>
</tr>
<tr>
<td>CONSTRUCTION_MAINTENANCE</td>
<td>-0.014</td>
<td>0.0018</td>
</tr>
<tr>
<td>RETAIL</td>
<td>-0.013</td>
<td>0.0001</td>
</tr>
<tr>
<td>AR_3</td>
<td>-0.013</td>
<td>0.0001</td>
</tr>
<tr>
<td>Skewness</td>
<td>-0.013</td>
<td>0.0001</td>
</tr>
<tr>
<td>Pay_bal_ratio</td>
<td>0.001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AR_6</td>
<td>0.001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>MEDICAL</td>
<td>0.02</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>CAR_RENTAL</td>
<td>0.023</td>
<td>0.0212</td>
</tr>
<tr>
<td>AUTOMOTIVE</td>
<td>0.024</td>
<td>0.0035</td>
</tr>
<tr>
<td>CASH_ADVANCE</td>
<td>0.024</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>FINANCIAL_SERVICES</td>
<td>0.037</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>kurtosis</td>
<td>0.06</td>
<td>0.0491</td>
</tr>
</tbody>
</table>

Positive coefficients are associated with increased risk and negative coefficients are associated with decreased risk.

The above Table shows that the Autoregressive parameters (AR_3, AR_5 and AR_6) extracted from the time series are amongst the significant variables in predicting future bad customers. This highlights the usefulness of model based parameter extraction from time series for purpose of detecting future bad customers. But it is difficult to determine how the frequency of purchases affects risk on the basis of these results. Skewness and kurtosis are also amongst the significant variables highlighting the usefulness of time
series moments for this purpose. These measures are extracted from the daily transaction amounts over the month for each customer. Pay_bal_ratio is the ratio of the payment to the amount of money owed by the customer at the start of the month. HEALTH_BEAUTY describes the percentage of amount of money spent for beauty or health products such as beauty parlors and gymnasium. EDUCATION represents the percentage of money spent on school, college and university education. CONSTRUCTION_MAINTENANCE represents the percentage of money spent on construction and maintenance such as building materials stores, paint and wallpaper stores and other similar stores. RETAIL represents the percentage of money spent on retail product such as florist, bookstores and newspaper stores. MEDICAL represents the percentage of money spent on medicine, doctors and medical procedures. CAR_RENTAL as the name suggests is the percentage of money spent on car renting. AUTOMOTIVE describes the percentage of money spent on automotive products such as Auto & Home Supply Stores, Motorcycle Dealers and Automobile and Truck Dealers. CASH ADVANCES are the percentage of amount of ATM cash withdrawals or cash withdrawn from banks. FINANCIAL SERVICES is one of the created “description codes” indicating percentage of money spend for banking services, credit union, money order and insurance.

Performance measures

The performance of a classification model is measured based on the classification accuracy of the model in the case of the test data. The performance depends on how well the model can correctly classify the bad customers as bad and good customers as good. The Percent Correctly Classified (PCC) is the most commonly used performance measure in the literature however most credit risk practitioners in financial institutions prefer the Gini coefficient as a measure of performance in the field of credit scoring. Both PCC and Gini coefficients are discussed below. Table 7.4 shows the confusion matrix for the test data using a cut off value of 0.5 for the probability of a bad future customer for stepwise logistic regression. The cut off value of 0.5 is the default option in SAS. This cut off value is very common although it may not be the most appropriate cut-off value. More appropriate cut off values are discussed later in this section.
Table 7.4  Confusion Matrix for Test Data Using Stepwise Logistic Regression

<table>
<thead>
<tr>
<th>Predicted</th>
<th>Actual</th>
<th>Good</th>
<th>Bad</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
<td>17119</td>
<td>3854</td>
<td>20973</td>
</tr>
<tr>
<td></td>
<td>82%</td>
<td>18%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td>1385</td>
<td>3751</td>
<td></td>
<td>5136</td>
</tr>
<tr>
<td></td>
<td>27%</td>
<td>73%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PCC=80%, cut-off probability=0.5

In scoring, the model performance is measured by how well the model can correctly classify the bad customers as bad and good customers as good. The above Table shows that 73% of the future bad customers are correctly classified by this model while only 18% of good customers are misclassified as bad. The PCC is 80% implying that 80% of the customers in the test data are correctly classified when a cut-off probability of 0.5 is used. This cut-off probability is misleading since it gives equal importance to misclassifying a bad customer as good and a good customer as bad. This point will be explained later in this chapter.

In credit scoring the cost of misclassifying bad customers as good customers is higher than misclassifying good customers as bad customers (Desai, et al., 1996). Overall the costs involved in misclassification are high (or very high); therefore the predictive accuracy of the classification model needs to be thoroughly evaluated. To reduce this cost we should aim to maximize the percentage of correctly classified bad customers (true negatives) whilst minimizing the number of good customers incorrectly classified as bad customers (false negative).

The following Table 7.5 shows the top percentiles for the predicted probabilities of future bad customers.
Table 7.5  
*Top Percentiles for the Estimated Probability of a Future Bad Customer*

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Future Good</th>
<th>Future Bad</th>
<th>% good customers misclassified as bad</th>
<th>% bad customers captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>30</td>
<td>231</td>
<td>0.1%</td>
<td>4.5%</td>
</tr>
<tr>
<td>2%</td>
<td>88</td>
<td>434</td>
<td>0.4%</td>
<td>8.5%</td>
</tr>
<tr>
<td>3%</td>
<td>173</td>
<td>610</td>
<td>0.8%</td>
<td>11.9%</td>
</tr>
<tr>
<td>4%</td>
<td>321</td>
<td>723</td>
<td>1.5%</td>
<td>14.1%</td>
</tr>
<tr>
<td>5%</td>
<td>428</td>
<td>876</td>
<td>2.0%</td>
<td>17.1%</td>
</tr>
<tr>
<td>6%</td>
<td>530</td>
<td>1035</td>
<td>2.5%</td>
<td>20.2%</td>
</tr>
<tr>
<td>7%</td>
<td>628</td>
<td>1199</td>
<td>3.0%</td>
<td>23.3%</td>
</tr>
<tr>
<td>8%</td>
<td>729</td>
<td>1363</td>
<td>3.5%</td>
<td>26.5%</td>
</tr>
<tr>
<td>9%</td>
<td>816</td>
<td>1531</td>
<td>3.9%</td>
<td>29.8%</td>
</tr>
<tr>
<td>10%</td>
<td>924</td>
<td>1693</td>
<td>4.4%</td>
<td>33.0%</td>
</tr>
</tbody>
</table>

Total good =20,983; Total Bad =5,126

In the above Table the top 10 percentiles with the highest predicted probabilities for future bad customer are shown with number of bad and good customers included in each percentile for the test data. The test data has 26,109 customers including 5126 future bad customers. The top 1% of the customers with highest predicted probabilities includes 231 future bad customers and only 30 future good customers. If we consider the top 1% of the customers as the threshold (cut-off), 4.5% of the 5126 future bad customers are correctly classified and 0.1% of the 20983 good customers are incorrectly classified as bad. If the top 5% of the customers with highest predicted probabilities are considered as the threshold, then 17.1% of all bad customers are correctly classified, while only 2% of all good customers are incorrectly classified as bad. Considering the top 10% of the customers with highest probability of future bad as the threshold then we have 33% of all bad customers correctly classified and 4.4% of all good customers incorrectly classified as bad customers.

To reduce the cost of misclassification, special attention should be given to the relative costs of the two types of misclassification errors. An appropriate threshold is selected based on the type of portfolio. We did not have access to the costs for these two misclassification errors and have therefore left the cut-off decision for the bank to make.
The actual classification results depend on the threshold to which the predicted probabilities for each outcome category are compared. The Receiver Operating Characteristics (ROC) curve, and the Gini coefficient, calculated using the area under the ROC curve provides measures of discrimination that are not dependent on arbitrary choices of threshold. These are the preferred performance measures. The ROC curve for the Stepwise Logistic Regression model is shown in Figure 7.2. The Gini coefficient for this model on the training dataset is 0.71 and for the test data it is 0.72. The Gini coefficients for the test data and the training data do not differ by more than 10% implying that the model is not over fitted. In behaviour scoring models a Gini coefficient of between 0.65 and 0.8 is regarded as a good quality classification model (refer Table 3.3). The above stepwise logistic model is therefore considered a good model for predicting the probability of future bad customers 4 months in advance.

Credit scoring systems are commonly used for large portfolios. This means that a slight improvement in a model’s accuracy can reduce the lender’s risk and translate into significant future savings. Developing an accurate credit scoring model helps banks and other credit institutions to protect themselves from customers with high default risk and hence minimizes loss while increasing profit.

As discussed in chapter four it has been found in the literature that a combination of techniques can improve the accuracy of the model (Lee & Jung, 1999/2000).
Our aim is to use this approach to develop a new hybrid model which results in higher accuracy. A hybrid SOM based classification model is proposed in the next section to predict the probability of future bad customers. The prediction accuracy of this model is compared with that of the stepwise logistic regression described above.

Model-2: Hybrid SOM based classification model

Real life credit card data is multidimensional and very large in size. The data used in this study is extremely imbalanced and therefore it becomes difficult to capture the characteristics of future bad customers. Classifiers which ignore the imbalanced class distribution problem lose their ability to predict the minority class correctly (Yen & Lee, 2009). To consider the imbalanced class distribution, additional to selecting an undersampled data, clustering can be useful as explained below. Customers within each cluster have similar transaction and payment behaviour. Modeling risk within each separate cluster helps identify the characteristics that can differentiate between similarly behaved good and bad customer. Modeling risk separately for clusters with the largest proportion of delinquents helps identify the particular characteristics of bad customers within each of these clusters. Therefore in this study the three clusters with the highest percentage of future bad customers are selected for modeling and a suitable classifier is determined for each of these clusters. Clearly the class distribution in these 3 clusters is less imbalanced than in the rest of the dataset. These three models are then used to predict the probability of a future bad customer for all the customers in the dataset. Therefore for all the customers we have three predicted probabilities indicating how likely a customer is to miss 3 or more payments in the future. These probabilities are obtained using separate classifiers for each of the 3 worst clusters. Later in this chapter we will explain how these predicted probabilities along with other input variables are inputted to the final two stage model to predict the probability of a future bad customer and also to predict the balance owed by the customer in 4-months time.

Step a) SOM to group customers based on their behaviour.

There are many ways of clustering. In the previous chapter, hierarchical clustering was used because of the great visual advantage and the relatively small sample size. In this
analysis, hierarchical clustering is not appropriate since our dataset is very large. Self Organizing Maps (SOM) are widely used for clustering large datasets, also mainly because of their great visual advantage. A ‘Kohonen self organizing map’ was the algorithm used in this study to group customers. With SOM the number of clusters is pre-specified. Customers are initially randomly assigned to each of the pre-specified clusters. When a new customer’s details are inputted into the SOM the customer is assigned to the cluster which is most similar to it based on data for all the input variables. The profile for this cluster and the neighbouring clusters are altered in order to reflect the characteristics of this new customer. As a result of this process clusters with more similarities are arranged closer to each other and dissimilar clusters are farthest from each other. All the variables were standardised so that variables with more variability do not dominate the clustering. To produce a SOM, the number of clusters should be pre-specified. We fist tried 12 and 24 clusters, then 36, 54 and 100 clusters. The dataset consisted of 35,855 customers including 20% future bad customers. The purpose of the SOM was to produce meaningful clusters with similarly behaved customers within each cluster. The goal was to find groups of similarly behaved customers who are more likely to go bad in future. The percentage of future bad customers was therefore calculated for each cluster using the pre-specified number of clusters (12, 24, 36, 54 and 100). The goal was to find a SOM with a high percentage of delinquent customers in only a few clusters. This would allow separate models to be developed for the unique behaviours of high risk customers in each of these clusters. Only with the 36 cluster SOM was this achieved. Three clusters with a high percentage of bad future customers were selected for modeling. Table 7.6 shows the percentages of good and bad customers in each of these three clusters.

Table 7.6  
Information on the 3 worst clusters produced by SOM with 36 clusters

<table>
<thead>
<tr>
<th>Cluster No</th>
<th>Future Good Customers</th>
<th>Future Bad Customer</th>
<th>% Future Bad Per Cluster</th>
<th>% Future Bad Captured</th>
<th>Cumulative % Bad Captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>8538</td>
<td>8307</td>
<td>49%</td>
<td>48%</td>
<td>48%</td>
</tr>
<tr>
<td>4</td>
<td>1535</td>
<td>1949</td>
<td>56%</td>
<td>11%</td>
<td>59%</td>
</tr>
<tr>
<td>21</td>
<td>1535</td>
<td>946</td>
<td>38%</td>
<td>5%</td>
<td>64%</td>
</tr>
<tr>
<td>all clusters</td>
<td>69624</td>
<td>17406</td>
<td>20%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>
The top three clusters - clusters with the highest percentage of future bad customers were selected for modeling. The cluster with the highest number of future bad customer was cluster 33. This cluster has 16,845 customers. This cluster includes 8,307 (49%) future bad customers. Therefore 48% of the total future bad customers are in cluster 33. The second worst cluster (cluster 4) includes 3,484 customers including 56% future bad customer. 11% of the total bad customers are in this cluster. Cluster 21 is a small cluster including 38% future bad customers and 62% future good customers. This cluster included 5% of the total future bad customers. Overall these three clusters include 64% of the total bad customers. There is sufficient information about the future bad customers in these three clusters to justify separate risk models for each of these clusters and clearly there is no problem with data imbalance in any of these clusters.

Each of these clusters is unique. Within each cluster customers have similar transaction and behaviour characteristics. On the other hand customers in different clusters have different transaction and behaviour characteristics. As said before the SOM not only groups similar customers in one cluster but also has the ability to group similar clusters close to each other. The top three clusters produced by SOM are not close to each other as shown in Figure 7.3, implying that the people in these three clusters do not behave similarly. Therefore modeling each of these dissimilar clusters separately results in discovering additional and unique characteristics of future bad customers.

<table>
<thead>
<tr>
<th>cluster1</th>
<th>cluster2</th>
<th>cluster3</th>
<th>cluster4</th>
<th>cluster5</th>
<th>cluster6</th>
<th>cluster7</th>
<th>cluster8</th>
<th>cluster9</th>
</tr>
</thead>
<tbody>
<tr>
<td>cluster10</td>
<td>cluster11</td>
<td>cluster12</td>
<td>cluster13</td>
<td>cluster14</td>
<td>cluster15</td>
<td>cluster16</td>
<td>cluster17</td>
<td>cluster18</td>
</tr>
<tr>
<td>cluster19</td>
<td>cluster20</td>
<td>cluster21</td>
<td>cluster22</td>
<td>cluster23</td>
<td>cluster24</td>
<td>cluster25</td>
<td>cluster26</td>
<td>cluster27</td>
</tr>
<tr>
<td>cluster28</td>
<td>cluster29</td>
<td>cluster30</td>
<td>cluster31</td>
<td>cluster32</td>
<td>cluster33</td>
<td>cluster34</td>
<td>cluster35</td>
<td>cluster36</td>
</tr>
</tbody>
</table>

[Figure 7.3 Clusters Produced by SOM]

Mean values for the worst three clusters for a few selected characteristics of are described below in Table 7.7.
Table 7.7  Characteristics of the Top Three Bad Clusters

<table>
<thead>
<tr>
<th>Cluster number</th>
<th>Mean</th>
<th>Skewness</th>
<th>Kurtosis</th>
<th>Pay spend ratio</th>
<th>Pay bal ratio</th>
<th>Financial services</th>
</tr>
</thead>
<tbody>
<tr>
<td>33</td>
<td>1.80</td>
<td>5.49</td>
<td>30.26</td>
<td>6.00%</td>
<td>15.41%</td>
<td>99.19%</td>
</tr>
<tr>
<td>4</td>
<td>5.80</td>
<td>5.48</td>
<td>30.27</td>
<td>1.34%</td>
<td>0.02%</td>
<td>97.11%</td>
</tr>
<tr>
<td>21</td>
<td>16.20</td>
<td>5.25</td>
<td>28.23</td>
<td>1.29%</td>
<td>0.02%</td>
<td>85.34%</td>
</tr>
</tbody>
</table>

All the three worst clusters have similar skewness values, all the three clusters have a high mean kurtosis value and all the customers in these three clusters, on average, spend more than 85% of their monthly expenditure on financial services (financial services include services such as government loans, government payment service, tax preparation services and credit union/money exchange service). However these clusters differ in terms of payment behaviour and other characteristics. As seen in the above Table payment to expenditure ratio (pay_bal_ratio) and balance to expenditure ratio (pay_bal_ratio) for cluster 33 is different to the other two bad clusters. Similarly Mean (average spending per day) is different across the three clusters. On average, customers in cluster 21 spend more money compared to customers in the other two clusters. Customers in cluster 21 on average spend around 16 dollars per day. Looking at the other characteristics it was observed that customers in cluster 21 spend money on health and beauty/ sport and recreation with a high percentage for cash advances (money withdrawn from an ATM using a credit card). However these customers’ payment behaviour seems to be regular. This is perhaps because these customers withdraw cash by credit card and repay the minimum amount of money each month so that they can use their credit card for some time before the bank recognizes them as bad customers. Customers in cluster 33 and 4 have a percentage of less than 1% for cash advances. All these points explain the diversity of the characteristics of the top three bad clusters. Capturing these differences will therefore assist in more accurate identification of bad customers. Hence each cluster was modeled separately to discover the unique characteristics of the bad customers within each of these three clusters.
**Step b) modeling for each of the top bad clusters:**

Customers in each of the three worst clusters were selected for modeling. The steps followed for each cluster are explained below.

1. Customers in a particular cluster are subdivided into training (40%), validation (30%) and test (30%) datasets.

3. Different classification techniques are applied to the training dataset to predict the probability of customers who are likely to miss 3 or more monthly payments within 4 months. The desired classifier is selected based on the gini coefficient for the test dataset. “Classifier_n” indicates the classifier chosen for cluster n.

4. Each “Classifier_n” is applied to the whole dataset producing an estimate for the probability of a customer becoming a future bad customer based on the n\textsuperscript{th} cluster behaviour. Let us call this probability “Predicted_n”, indicating the estimated probability of a future bad customer based on cluster n risk behaviour.

All the above steps are repeated for the top three bad clusters producing three predicted probabilities of future bad customers based on the different cluster behaviours.

As mentioned in Table 7.6 the top three bad clusters are cluster 33, 4 and 21. Cluster 33 includes 16,845 customers with 51% future good customers and 49% future bad customers.

Classification Trees, logistic regression and neural networks were applied to this cluster to predict the probability of future bad customers. Table 7.8 shows Gini coefficients for these different techniques.

<table>
<thead>
<tr>
<th>Classification technique</th>
<th>Gini Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Backward logistic regression</td>
<td>0.65</td>
</tr>
<tr>
<td>Stepwise logistic regression</td>
<td>0.65</td>
</tr>
<tr>
<td>Classification tree (gini)</td>
<td><strong>0.76</strong></td>
</tr>
<tr>
<td>Classification (entropy)</td>
<td>0.76</td>
</tr>
<tr>
<td>2 hidden node NN*</td>
<td>0.68</td>
</tr>
</tbody>
</table>

\textit{NN*}: variable are selected using stepwise logistic regression since there is no inbuilt variable selection option available for NN
Table 7.8 and the ROC curves in Figure 7.4 show that the Classification tree using gini reduction and entropy criterion have the minimum misclassification rate. The significant variables using this model are shown in Table 7.9. The time series parameter AR (1) is amongst these variables indicating the usefulness of time series clustering. The whole data set is scored based on this model to predict the Probability of a customer becoming a bad customer within 4 months and the predicted probabilities are denoted as 'predicted_33'.

The same procedure is applied to the customers in clusters 21 and 4. The best classifier for cluster 21 was a classification tree using the gini reduction criterion. For cluster 4 the stepwise logistic regression had the highest Gini coefficient. The models were applied to all customers in the sample data and probabilities of future bad customers based on cluster 4 and 21 were also predicted for all customers (predicted_4 and predicted_21). The significant variables for predicting future bad customers for cluster 4, 21 and 33 data are given in Table 7.9.
### Table 7.9  The Significant Variables Based on Cluster 4, 21 and 33

<table>
<thead>
<tr>
<th>Cluster 4</th>
<th>Cluster 21</th>
<th>Cluster 33</th>
</tr>
</thead>
<tbody>
<tr>
<td>FINANCIAL SERVICES</td>
<td>FINANCIAL SERVICES</td>
<td>FINANCIAL SERVICES</td>
</tr>
<tr>
<td>Pay_Spend_Ratio</td>
<td>Mean</td>
<td>Pay_Spend_Ratio</td>
</tr>
<tr>
<td>Std</td>
<td>Pay_Bal_Ratio</td>
<td>Mean</td>
</tr>
<tr>
<td>Ar_1</td>
<td>End_start_bal_ratio</td>
<td>Pay_Bal_Ratio</td>
</tr>
<tr>
<td>Sk</td>
<td></td>
<td>End_start_bal_ratio</td>
</tr>
<tr>
<td>Cash_advance</td>
<td></td>
<td></td>
</tr>
<tr>
<td>End_start_bal_ratio</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

In cluster 33, customers who spend more than 93% on financial services and have a low ratio value for balance at the end of the month to the balance at start of the month and also have low mean expenditure, are more risky compared to those with higher mean expenditure or higher ratio of change of balance at start and end of the month. Also in this cluster customers who spend less than 93% on financial services but have low payment to expenditure ratio (pay_spend_ratio) are more risky compared to customers who have high ratio of payment to expenditure ratio.

In cluster 21, customers who have a smaller ratio for end of month balance to start of the month balance (end_start_bal_ratio) and for whom average daily expenditure (mean) is low are more risky, i.e., more likely to become future bad customers. Also in this cluster customers who have high ratio of end of month balance to start of the month balance (end_start_bal_ratio) and spend a high percentage of their expenditure on financial services are more risky compared to customers who spend a lower percentage on financial services. Also customers with high values for end balance to start balance ratio and whose payment to balance ratio is very low are risky customers. These rules were found using customers in cluster 22 which will be applied to all other customers in the next step (step c).

In cluster 4, customers with smaller end to start of the month balance ratio are less risky compared to customers with higher end to start balance ratio. Also amongst the
customers with high ratio of end to start balance ratio, customers who spend more than 97% on financial services are highly risky and also customers with smaller AR 1 (AR coefficient with lag one) coefficient are more risky compared to those with larger phi1.

Modeling based on these three different clusters, captures the distinct characteristics of the bad customers within each of these clusters. Each of these models/rules produces different significant variables for predicting future bad customers. These rules distinguish between high risk and low risk customers with very similar spending patterns. In other words these rules identify the differences “within” the clusters; considering these rules results in more accurate prediction of future bad customers. However these rules are based on only the top three clusters and do not distinguish between high risk and low risk customers outside these clusters. We also need to identify the characteristics which can identify high risk and low risk customers outside of these clusters.

To do this, the probabilities produced from classifier 4, 21 and 33 for all customers, along with other input variables given in Table 6.10 are then inputted to the final two stage model to predict the probability of a future bad customer in 4 months time, and also the amount owed by a customer to the bank at that time.

**Step c) Final model to predict future performance based on the input variables and the estimated probabilities for a future bad customer based on the top three clusters:**

Predicted probabilities based on the top three bad clusters were added to the other existing input variables to form 48 input variables for the final modeling of the future risk using the sample dataset with 20% future bad customers.

As for each of the cluster samples the complete sample dataset is partitioning into a training dataset (40%), a validation dataset (30%) and a test dataset (30%).

As before using appropriate variable selection methods different techniques, namely Neural Networks, Classification Tree and Logistic Regressions were applied to predict the probability of future good/bad customers. Table 7.10 shows the misclassification rate for the different techniques based on a cutoff of 0.5 for the estimated probability of a future bad customer.
The above Table shows that the classification trees and Neural Network with 5 hidden nodes have the lowest misclassification rate. Freeman and Moisen (2008) conclude in their discussion that PCC (1-misclassification rate) can be misleading when data is unbalanced and therefore this measure of performance has little value in practice. In the current classification literature, many studies still use this measure for comparing the performance of classifiers, perhaps because of its simplicity. At this stage by comparing the performance of the techniques using ROC curve we are going to prove how the PCC can be misleading.

Figure 7.5 shows the ROC curve for stepwise logistic regression, a classification tree using the gini criterion and neural networks with 2 hidden nodes and 5 hidden nodes. Forward logistic regression showed exactly the same ROC curve as stepwise logistic regression. NN with 5 hidden node and 3 hidden nodes also had the same ROC curves.

<table>
<thead>
<tr>
<th>Classification technique</th>
<th>Training set</th>
<th>Validation set</th>
<th>Test set</th>
</tr>
</thead>
<tbody>
<tr>
<td>Forward logistic regression</td>
<td>0.089</td>
<td>0.088</td>
<td>0.087</td>
</tr>
<tr>
<td>Stepwise logistic regression</td>
<td>0.089</td>
<td>0.088</td>
<td>0.087</td>
</tr>
<tr>
<td>Classification tree (gini)</td>
<td>0.081</td>
<td>0.080</td>
<td>0.080</td>
</tr>
<tr>
<td>Classification tree (entropy)</td>
<td>0.081</td>
<td>0.082</td>
<td>0.080</td>
</tr>
<tr>
<td>2 hidden node NN*</td>
<td>0.085</td>
<td>0.083</td>
<td>0.082</td>
</tr>
<tr>
<td>3 hidden node NN*</td>
<td>0.082</td>
<td>0.082</td>
<td>0.081</td>
</tr>
<tr>
<td>5 hidden node NN*</td>
<td>0.083</td>
<td>0.081</td>
<td>0.080</td>
</tr>
</tbody>
</table>
Figure 7.5  ROC- Curve for different techniques used on the validation dataset.

The above Figure shows that the classification tree performs worse than the other methods. The ROC curve for Logistic regression and neural networks intersect suggesting that the performance of these two techniques is almost the same. The Gini values for these methods are shown in the Table 7.11.

<table>
<thead>
<tr>
<th>Classification Technique</th>
<th>Gini Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN(5hn)</td>
<td>0.882</td>
</tr>
<tr>
<td>NN(2hn)</td>
<td>0.873</td>
</tr>
<tr>
<td>Logistic regression(stepwise)</td>
<td>0.878</td>
</tr>
<tr>
<td>Tree(gini)</td>
<td>0.748</td>
</tr>
</tbody>
</table>

The Gini coefficient for the gini classification tree is the lowest amongst the above techniques. The NN with 2 hidden nodes is next and the stepwise logistic regression is a
little better. The NN with 5 hidden nodes performs slightly better. However, a NN is not easily interpretable and it has a longer execution time. Logistic regression is easily interpretable and well understood. Therefore stepwise logistic regression is selected as the final classification model in the proposed hybrid SOM based classification model.

The significant variables for the prediction of future bad customers using the hybrid SOM based classification model are given in Table 7.12.

Table 7.12  
Significant Variables Using Hybrid SOM based classification Model

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>intercept</td>
<td>-5.720</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AR_5</td>
<td>-0.740</td>
<td>0.005</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>-0.020</td>
<td>0.025</td>
</tr>
<tr>
<td>HEALTH_BEAUTY</td>
<td>-0.017</td>
<td>0.024</td>
</tr>
<tr>
<td>MEDICAL</td>
<td>-0.016</td>
<td>0.007</td>
</tr>
<tr>
<td>CLOTHING</td>
<td>-0.011</td>
<td>0.016</td>
</tr>
<tr>
<td>RETAIL</td>
<td>-0.007</td>
<td>0.007</td>
</tr>
<tr>
<td>AUTOMOTIVE</td>
<td>-0.006</td>
<td>0.025</td>
</tr>
<tr>
<td>End_start_bal_ratio</td>
<td>0.0001</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Predicted_4</td>
<td>0.016</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>CASH_ADVANCE</td>
<td>0.017</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>KR</td>
<td>0.023</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>FINANCIAL_SERVICES</td>
<td>0.023</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AR_3</td>
<td>0.770</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Predicted_21</td>
<td>1.419</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Predicted_33</td>
<td>4.610</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>

Positive coefficients are associated with increased risk and negative coefficients are associated with decreased risk.

AR_3 and AR_5 are the autoregressive parameters of lag 3 and lag 5 extracted from the daily transaction time series data for each customer. The significance of these variables highlights the usefulness of model based parameter extraction from time series. However, it is difficult to determine the relationship between the frequency of shopping and financial risk based on these variables. Predicted probabilities of future risk based on cluster 4, 21 and 33 are amongst the significant variables which show the usefulness of cluster based classification. KR is the kurtosis for the amount of transactions over the
month extracted from daily transaction time series. End_start_bal_ratio is the ratio of end of month balance to the balance at the start of the month. EDUCATION represents the amount of money spends on school, college and university educations. HEALTH_BEAUTY describes the percentage of amount of money spends for beauty or health products such as beauty parlors and gymnasium. MEDICAL represents a description code which relates to money spent on medicine, doctors and medical procedures. CLOTHING represents a description code which relates to money spent on clothing Stores, accessories, shoe stores and other related products. RETAIL is a description code describing the percentage of amount of money spends on retail product such as florist, bookstores and newspaper stores. AUTOMOTIVE describes money spent on automotive products such as Auto & Home Supply Stores, Motorcycle Dealers and Automobile and Truck Dealers. END_START_BAL_RATIO is the ratio of outstanding balance to current month’s expenditure. CASH ADVANCES are the ATM cash withdrawals or cash withdrawn from banks. These extra variables identify the behaviours associated with the behaviour of bad customers outside of the top 3 bad clusters. FINANCIAL SERVICES is one of the created “description codes” indicating money spent on government services, government loans, tax preparation services and credit union/ money order services. The results in Table 7.12 indicate that the risk is lower for customers who spend more on automotive, retail, clothing, medical, health and beauty products or services and education.

Performance Measure:
The Gini coefficient and confusion matrix for the test data using the Hybrid SOM based Classification Model is calculated. Table 7.13 shows the confusion matrix using a cut-off value of 0.5.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Good</th>
<th>Bad</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td>20464</td>
<td>509</td>
<td>20973</td>
</tr>
<tr>
<td></td>
<td>97.6%</td>
<td>2.4%</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td>1667</td>
<td>3469</td>
<td>5136</td>
</tr>
<tr>
<td></td>
<td>32.5%</td>
<td>67.5%</td>
<td></td>
</tr>
</tbody>
</table>

PCC=92%, cut-off probability=0.5
Using the Hybrid SOM based Classification Model with a cutoff value of 0.5, almost 68% of the bad customers are correctly classified and only 2.4% of the good customers are incorrectly classified. However if the cut-off value is adjusted according to predicted probabilities the models’ accuracy increases. Table 7.14 shows a range of thresholds.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Future Good</th>
<th>Future Bad</th>
<th>% good customers misclassified as bad</th>
<th>% bad customers captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>1</td>
<td>260</td>
<td>0.00%</td>
<td>5%</td>
</tr>
<tr>
<td>2%</td>
<td>1</td>
<td>521</td>
<td>0.00%</td>
<td>10%</td>
</tr>
<tr>
<td>3%</td>
<td>1</td>
<td>782</td>
<td>0.00%</td>
<td>15%</td>
</tr>
<tr>
<td>4%</td>
<td>1</td>
<td>1043</td>
<td>0.00%</td>
<td>20%</td>
</tr>
<tr>
<td>5%</td>
<td>1</td>
<td>1304</td>
<td>0.00%</td>
<td>25%</td>
</tr>
<tr>
<td>6%</td>
<td>1</td>
<td>1565</td>
<td>0.00%</td>
<td>30%</td>
</tr>
<tr>
<td>7%</td>
<td>1</td>
<td>1826</td>
<td>0.00%</td>
<td>36%</td>
</tr>
<tr>
<td>8%</td>
<td>11</td>
<td>2077</td>
<td>0.05%</td>
<td>40%</td>
</tr>
<tr>
<td>9%</td>
<td>13</td>
<td>2336</td>
<td>0.06%</td>
<td>45%</td>
</tr>
<tr>
<td>10%</td>
<td>29</td>
<td>2582</td>
<td>0.14%</td>
<td>50%</td>
</tr>
</tbody>
</table>

Total good = 20,983; Total Bad = 5,126

In the above Table the top 10 percentiles with highest predicted probabilities are shown with the number of bad and good customers included in each percentile. The test data has 26,109 customers including 5,126 future bad customers. The top 1% of the customers with highest predicted probabilities includes 260 future bad customers and only 1 future good customer. If we consider the top 1% of the customers as a threshold, 5% of bad customers are correctly classified and only 1 good customer is incorrectly classified as bad. If the threshold is set to the top 5% of the customers with the highest predicted probabilities, then 1,304 customers are considered as future bad customers. Amongst these predicted future bad customers only 1 customer is a good customer classified as bad. 1,304 future bad customers were predicted correctly. These correctly classified bad customers make up 25% of all bad customers. Therefore we capture 25% of the future bad customers in the top 5% of the customers with highest predicted probabilities. The top 10% of the customers with highest predicted probabilities include
2611 customers. Considering the top 10% of customers as a threshold we correctly classify 2582 future bad customers and only 29 future good customers are incorrectly classified as bad. Considering the top 10% of the customers with highest predicted probabilities as a threshold we have captured 50% of all future bad customers.

The ROC curve for the Hybrid SOM based classification model is shown in Figure 7.6. The area under the curve can be viewed as the proportion of correct classification across all the possible thresholds.

![ROC Curve](image)

Gini Coefficient = 0.87

Figure 7.6 ROC Curve for Hybrid SOM based Classification Model using test data.

The Gini coefficient for the training dataset is 0.86 and for the test dataset it is 0.87. The evaluation Table for Gini coefficient in chapter 3 (Table 3.3) shows that a Gini coefficient of greater than 0.80 indicates very good classification quality. The proposed Hybrid SOM-based Classification Model is therefore a very good model for behaviour scoring.

**Comparison of model -1 and model-2 classification results:**

As discussed in chapter 2 the purpose of combining models is to produce a final model that is better than the individual models in terms of classification accuracy rate or other criteria (Lee & Jung, 1999/2000). Combining models often results in the improvement
of a model’s prediction accuracy compared to that of the individual models’ performances. However, building a combined model is not always easy. It is more time consuming to develop and utilize combined models than individual models and it is more difficult to interpret the rules generated by the combined model.

Classification accuracy is not the only concern for credit scoring analysts when building score cards. Other aspects such as speed of classification and interpretability should also be considered. The extra time and complicated rules required for combined models may not be reasonable if the improvement is minimal.

Model-1 and Model-2 are compared using ROC curves and their Gini coefficients. The ROC curves for both the models are shown in Figure 7.7. The more an ROC curve approaches the point (0,1), the better is the classifier.

![ROC Curve](image)

Gini Coefficient using Model-2 = 0.87
Gini Coefficient using Model-1 = 0.72

**Figure 7.7** ROC curves for hybrid SOM based & stepwise LR models.

From the above Figure we can conclude that model-2 (hybrid SOM based classification model) is much better than model-1 (stepwise logistic regression). This is confirmed by Gini coefficients. The Gini coefficient for the test data using stepwise logistic regression is 0.72. The Gini coefficient for the test dataset using hybrid SOM based classification model is 0.87. The hybrid SOM based Classification Model is superior not only in
correctly classifying bad customers (true negatives) but also this model has a lower percentage of incorrectly classified good customers (false negative). This is because we have grouped the customers into smaller clusters and modeled the bad clusters before the final prediction. Prediction on each of the three bad clusters is significantly increasing the accuracy of the final model. Table 7.15 summarizes the top 10% of customers with highest predicted probabilities using Stepwise Logistic Regression (Model-1) and Hybrid SOM based Classification Model (Model-2). As shown in the Table the top 10% of customers with the highest predicted probabilities using model-1 include 1693 future bad customers and 924 future good customers. Using model-2 the top 10% of customers with highest predicted probabilities includes 2,582 future bad customers and only 29 future good customers.

Table 7.15  
Top 10% of customers with highest predicted future bad probabilities

<table>
<thead>
<tr>
<th>Models</th>
<th>Future good customer</th>
<th>Future Bad customer</th>
<th>% Future Good Classified as Bad</th>
<th>% Future Bad Captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Top 10%</td>
<td>Model-1</td>
<td>924</td>
<td>1693</td>
<td>4.4%</td>
</tr>
<tr>
<td></td>
<td>Model-2</td>
<td>29</td>
<td>2582</td>
<td>0.14%</td>
</tr>
</tbody>
</table>

Total good =20,983; Total Bad =5,126

If the top 10% of the predicted future bad probabilities is used as the threshold to decide future bad customers, using model-1 we have correctly classified 33% of the future bad customers and 4.4% of the future good customers are incorrectly classified as future bad customers. Using model-2 50% of the future bad customers are correctly classified and almost no good customer are incorrectly classified as future bad customers.

The main purpose of Behaviour scoring models is to manage the risk of existing customers. Based on the scores produced by the model banks make decisions to increase/decrease interest rates and increase/ decrease credit limits and also banks come up with strategies to deal with risky customers. Therefore it is very important that the model correctly classifies the risky customers’ as well as the good customers.

Existing very good customers should be predicted as lowest future risks (future good customers) so that the bank makes the right decisions when increasing credit limits (of good customers). Classifying bad customers as low risk (future good customers) may become very costly for banks if high credit limits are set for these customers. If a model
classifies future risky customers as future good customers then banks may increase the credit limits of these customers resulting in serious losses for the bank. Therefore in behaviour scoring it is not only important to classify the future bad customers accurately but it is also important to accurately classify very good customers. This is another reason why PCC should not be used as a performance measure in behaviour scoring models. Table 7.16 summarizes the bottom 10% of customers with lowest predicted risk probabilities using Stepwise Logistic Regression (Model-1) and the Hybrid SOM based Classification Model (Model-2).

Table 7.16  

<table>
<thead>
<tr>
<th>Models</th>
<th>Future good</th>
<th>Future Bad</th>
<th>% future bad classified as good</th>
<th>% future Good captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>Bottom 10%</td>
<td>Model-1</td>
<td>2848</td>
<td>24</td>
<td>0.47%</td>
</tr>
<tr>
<td>10%</td>
<td>Model-2</td>
<td>2859</td>
<td>13</td>
<td>0.25%</td>
</tr>
</tbody>
</table>

Total good =20,983; Total Bad =5,126

The bottom 10% (or more) of the customers with lowest predicted probabilities of future risk of going bad (higher credit scores) would often be granted a higher credit limit. Both the models perform well in capturing good customers in the bottom 10% of the data with the lowest predicted probabilities for future risk. In the bottom 10% of customers with lowest predicted probabilities there are only 13 future bad customers using model-2 and 24 future bad customers when model-1 is used. However, as seen previously the same is not true when it comes to the capture of bad customers in the top 10% of the data.

The stepwise logistic regression model is considered to be a good classification model for predicting the probability of future bad customers 4 months in advance; however, the proposed Hybrid SOM-based classification Model is a much better classification model.
7.4.2 Scoring the developed models to highly imbalanced datasets.

In the previous section the sample datasets used for modeling included 20% future bad customers. However, the entire dataset includes only 1.73% future bad customers. Both the models developed in the previous section are applied to the entire dataset from March 2007 to see if having a much smaller proportion of bad customers would decrease the accuracy of the models. Details of the complete March 2007 dataset is provided in Table 7.17.

Table 7.17  Target Definition and Sample Size for March 2007 Dataset

The sample used to train the model in the previous section is a stratified sample from the complete March 2007 dataset. Under-sampling is performed so that the rare class becomes less rare in the sample, therefore reducing the effect of imbalance and increasing the accuracy of the model predictions. In the sample dataset the majority class is under sampled to get 20% future bad and 80% future good customers in the dataset. Therefore the proportion of future bad customers in the training dataset differs significantly from this proportion in the population. The correct priors should be specified in order to adjust the classification boundary (cutoff) when scoring new data. This can be done after the model is applied to the new dataset. A simple formula is used to adjust the posterior probabilities for priors given as

\[
\text{Posterior}_{1}(\text{new}) = \frac{\text{Posterior}_{1}(\text{old}) \times \text{prior}_{1}(\text{new}) / \text{prior}_{1}(\text{old})}{\sum_{j} \text{Posterior}_{j}(\text{old}) \times \text{prior}_{j}(\text{new}) / \text{prior}_{j}(\text{old})}
\]
where j= 0, 1 representing good and bad customers respectively.
Prior(old)= 80 % or 20% for good and bad customers respectively.
Prior (new)= 99.3% or 1.7% for good and bad customers respectively.
Posterior(old)= predicted probability based on the original 20:80 model.

However SAS Eminer incorporates this calculation if the good bad odds of both the dataset (one which model is built on and one which model is applied to) are provided before scoring the model.

After doing this, both the models, Stepwise Logistic Regression and Hybrid SOM Based Classification Model were applied to the complete March 2007 dataset.

Model-1: Stepwise Logistic regression developed based on March data:
The confusion matrix for the entire dataset using the stepwise logistic model is shown in Table 7.18 when a cut off value of 0.5 is selected.

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th>Good</th>
<th>Bad</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Good</td>
<td></td>
<td>883045</td>
<td>103064</td>
<td>986109</td>
</tr>
<tr>
<td></td>
<td>90%</td>
<td>10%</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td></td>
<td>8464</td>
<td>8942</td>
<td>17406</td>
</tr>
<tr>
<td></td>
<td>49%</td>
<td>51%</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

PCC=89% , cut-off probability=0.5

The cut off value of 0.5 is not appropriate here since it leads to classifying many good customers as bad (see Table 7.18). If the 10,000 customers with the highest predicted probabilities of future risk of going bad are classified as bad customers almost 2000 bad customers are correctly identified and 8,000 good customers are incorrectly identified. Depending on what is more appropriate for the bank, the cut-off value is adjusted. The Stepwise Logistic Regression model produced a gini coefficient of 0.71 for the entire population.
Model-2: Hybrid SOM based classification model developed based on March data:

The Hybrid SOM based classification model was applied to the entire dataset including only 1.73% future bad customers. This classifier produced a Gini coefficient of 0.86. The confusion matrix for the entire dataset using this model is shown in Table 7.19 when a cut off value of 0.5 is selected.

Table 7.19  Confusion matrix for entire dataset using hybrid SOM based classification model

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
<td>Bad</td>
<td>Total</td>
</tr>
<tr>
<td>Good</td>
<td>981950</td>
<td>4159</td>
<td>986,109</td>
</tr>
<tr>
<td></td>
<td>99.6%</td>
<td>0.4%</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td>8319</td>
<td>9087</td>
<td>17,406</td>
</tr>
<tr>
<td></td>
<td>47.8%</td>
<td>52.2%</td>
<td></td>
</tr>
</tbody>
</table>

PCC=98% , cut-off probability=0.5

Using a hybrid SOM based classification model, not only is the percentage of bad customers correctly classified increased but also the percentage of good customers incorrectly classified as bad is decreased to 0.4%.

By classifying the 10,000 customers with the highest predicted probabilities for future bad customer, 8,800 bad customers were correctly identified and only 1,200 good customers were incorrectly classified as bad. By classifying the 20,000 customers with the highest predicted probabilities for future bad customer, 9,800 bad customers were correctly identified and only 9,200 good customers were incorrectly classified as bad.
Model-1 and Model-2 are compared using ROC curves and Gini coefficients.

![ROC Curve](image)

Gini Coefficient using Model-2 = 0.86
Gini Coefficient using Model-1 = 0.71

*Figure 7.6. ROC curves for hybrid SOM based classification model & stepwise LR.*

The ROC curves for both the models are shown in Figure 7.6. It is clearly shown that the hybrid SOM-based classification model is performing better than Stepwise logistic regression. This is confirmed by the Gini coefficients. The Gini coefficient for hybrid SOM based classification model is 0.86 and for the Stepwise logistic regression it is 0.71. Clearly this method can be used even when the proportion of bad customers is very small.

### 7.4.3 Future prediction based on the proposed model (Stability):

Results from the previous section may be biased since most of the bad customers were used to train the model. In this section we apply the models to a new dataset to see if the model derived from March 2007 data can predict the probability of customers missing 3 or more payments in August 2007 when applied to April 2007 data.

The previous study was performed on transaction data for customers in March 2007 to predict the probability of future performance within 4 months i.e., July 2007. In this section we monitor April transactional dataset to predict the probability of customers’ future performance using the models developed based on March 2007 data.
The same variables derived in section 6.3 are prepared for the April data. Target and sample sizes are shown in Table 7.20.

Table 7.20  

<table>
<thead>
<tr>
<th>Target Definition and Sample Size for April 2007 Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td>Aim: to predict the probability of future performance within 4 months for April 2007 transaction data using models developed based on March data.</td>
</tr>
<tr>
<td>Data Size 999,625 including 14,441(1.4%) future bad customers.</td>
</tr>
</tbody>
</table>
| Target = \[
\begin{cases} 
1 \text{ (bad)} & \text{if a customer is 3 or more payments overdue in August 2007} \\
0 \text{ (good)} & \text{if a customer is not overdue in August 2007}
\end{cases}
\] |

**Model-1: Stepwise Logistic regression developed based on March data:**
Applying the stepwise logistic model developed using the March 2007 data to the April dataset we estimate the probability of a customer missing 3 or more payments in the next 4 months. The Gini coefficient for the ROC curve in Figure 7.7 shows a slight decline to 0.68. This Gini coefficient still indicates a good model.
However, the top 10,000 customers with the highest predicted probability of becoming future bad customers include 9,000 good customers and only 1,000 bad customers. The top 20,000 customers with the highest predicted probability of becoming future bad customers include 18,200 good customers and only 1,800 bad customers.

**Model-2: Hybrid SOM based classification model developed based on March data**
Applying the Hybrid SOM classification model to the April dataset to produce probability estimates of future performance showed no deterioration, with a Gini coefficient still equal to 0.86. The ROC curve is shown in Figure 7.7
The top 10,000 customers with the highest predicted probability of becoming future bad customers include 3,000 good customers and 7,000 bad customers. The top 20,000 customers with the highest predicted probability of becoming future bad customers include 11,580 good customers and 8,420 bad customers.

The model developed using a sample of March 2007 customers resulted in good classification for April customers. Also this model was applied to May and June 2007 and the results were nearly as good.

Hence we can conclude that the Hybrid SOM base Classification Model developed using the March 2007 sample is time independent and the rules can be applied to other months to capture the financial situation of customers and future risk performance of the customers. However customers’ financial situations may change in December due to Christmas and holiday season. This model may not be appropriate for December since it is influenced by other factors. Perhaps a separate model should be developed for that particular month. Comparing with the stepwise logistic regression model, this model is superior not only in correctly classifying bad customers but also this model has a much lower percentage of incorrectly classified bad customers. However, both models perform well in incorrectly classifying good customers.
7.4.4 Prediction of Amount Owed:

Predicting the future performance is very important for banks to differentiate between good and bad customers. However, a customer missing 3 payments owing $500 and a customer missing 3 payments owing $10,000 are both considered as bad customers. Predicting the amount owed by customers is useful in differentiating between high risk future bad customers and low risk future bad customers.

In the proposed hybrid SOM based classification model, after finalizing the variables selected using stepwise logistic regression, instead of predicting only the probability of future risk, we use a two stage model to predict the probability of future risk at the first stage and the amount owed at the second stage. A two stage model includes two target variables. At the first stage there is a binary target and at the second stage a metric target is predicted. At the second stage the balance owed by customers in 4 months time is calculated using regression. Then the expected loss is calculated by multiplying the probability of a future bad customer calculated in the first stage by the predicted balance owed in the second stage. A higher expected loss indicates a very risky customer with much debt.

The balanced owed was highly skewed. Before applying regression a transformation was applied to maximize the normality. Square root transformation was the most appropriate transformation. Also we had to omit a few customers before analysis since customers who exceed 5 missed payments have no future balance owed because they have been moved to collections. However this affects the models’ prediction accuracy.

The ROC curve was drawn for the expected loss and the corresponding gini coefficient was found to be 0.805 which confirms that the model is good.

Incorporating the predicted amount owed by customers into the model is useful because it identifies for the bank the very risky customers who owe a large amount of money as opposed to the risky customers who owe a smaller amount.
The average predicted balance and actual balance for the top ten expected loss percentiles are provided in Table 7.21.

Table 7.21  \textit{Actual & Predicted balances for the top 10 percentiles for expected loss}

<table>
<thead>
<tr>
<th>Top percentiles</th>
<th>Actual cumulative mean balance</th>
<th>Predicted cumulative mean balance</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>14,634</td>
<td>13,603</td>
</tr>
<tr>
<td>2</td>
<td>20,627</td>
<td>19,761</td>
</tr>
<tr>
<td>3</td>
<td>25,892</td>
<td>25,037</td>
</tr>
<tr>
<td>4</td>
<td>31,137</td>
<td>30,314</td>
</tr>
<tr>
<td>5</td>
<td>40,388</td>
<td>39,008</td>
</tr>
<tr>
<td>6</td>
<td>48,200</td>
<td>46,346</td>
</tr>
<tr>
<td>7</td>
<td>54,987</td>
<td>52,687</td>
</tr>
<tr>
<td>8</td>
<td>60,997</td>
<td>58,267</td>
</tr>
<tr>
<td>9</td>
<td>67,365</td>
<td>64,101</td>
</tr>
<tr>
<td>10</td>
<td>76,841</td>
<td>72,594</td>
</tr>
</tbody>
</table>

In all the previous sections performance of customers was predicted 4 months in advance based on transaction and payment behaviour over one month. In the next
section the same methodology is applied to verify if the same variables can predict the behavior of customers 6 months in advance.

In previous sections models were developed using a March dataset to predict the future risk of going bad in the next 4 months. At this stage we wanted to see if the same procedure can be used to predict the future risk of customers 6 months ahead. Next section describes the model developed to predict the future risk of customers going bad 6 months in advance.

7.5 A 6-month time horizon:

Similarly to the previous analysis, March 2007’s transaction and payment datasets are selected for developing the model. However in this analysis a 6 month time horizon is considered. If a customer is not overdue by September 2007 then the customer is considered to be a good customer (target=0). If a customer has missed 3 or more payments by September 2007 then the customer is considered to be a bad customer (target=1). Similarly to the previous analysis inactive customers and customers who were 1 or 2 payments overdue by September 2007 were excluded from the analysis, since including these indeterminate customers affects the accuracy of prediction (refer 6.2.6). As before, customers with missing delinquency status were eliminated from the analysis. The same variables derived in chapter 6 section 6.3 are used as input variables to predict the future risk of customers 6 months in advance. Target and sample sizes are shown in Table 7.22

Table 7.22  Target Definition and Sample Size for April 2007 Dataset

<table>
<thead>
<tr>
<th>Aim: predict the probability of future performance 6 months in advance using March 2007 data Data Size 964,958 including 14,091(1.5%) future bad customers.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Target (future performance)</td>
</tr>
<tr>
<td>1 (bad) if a customer is 3 or more payments overdue in September 2007</td>
</tr>
<tr>
<td>0 (good) if a customer is not overdue in September 2007</td>
</tr>
</tbody>
</table>
The transaction and payment behavior of customers during March 2007 was monitored and the delinquency status of all these customers was checked for September 2007. There were 964,958 customers who matched the above defined target including 14,091 future bad customers. This dataset contains 51 input variables including customer id (see Table 6.10) and a binary target variable to identify each customer as a future good or bad customer.

This dataset is imbalanced with a very low proportion of future bad customers. Modeling based on this imbalanced data with only 1.5% future bad customers would classify all customers as future good customers and result in a high accuracy rate only for good customers. To get reliable results, a sample with approximately 20% future bad customers (target=1) and 80% future good customers was selected. In other words under-sampling was performed selecting all the future bad customers and randomly selecting future good customers so that the proportion of future bad to future good customers is 1:4. Under-sampling resulted in a sample of size 70,455 with 20% future bad customers.

7.4.1 Model development and validation

A random sample of data from March 2007 with 20% future bad customer is selected for analysis using the proposed Hybrid SOM based Classification Model to predict the probability of future risk 6 months in advance. This model is then compared to the previous Hybrid SOM based classification model based on a 4 month time horizon to see if the significant variables in predicting future risk remain the same.

Step a) SOM to group customers based on their behaviour.

All the variables were standardized so that variables with more variability do not dominate the clustering. As before, a ‘Kohonen self organizing map’ was used to group customers. Different number of clusters was specified. The goal was to find groups of similarly behaved customers who are more likely to go bad in the future. The percentage of future bad customers was calculated for each cluster using a pre-specified number of clusters (12, 24, 36, 54 and 100). The goal was to find a SOM with a high
percentage of delinquent customers in only a few clusters. As before, only with the 36 cluster SOM was achieved. In this study only two clusters with a high percentage of bad future customers were selected for modeling. Besides these two clusters, there was no other cluster with a high percentage of future bad customers that could be modeled on its own. Table 7.23 shows the percentages of good and bad customers in each of these clusters.

Table 7.23  Information on the 36 clusters produced by SOM

<table>
<thead>
<tr>
<th>Cluster No</th>
<th>Future Good</th>
<th>Future Bad</th>
<th>% Future Bad Per Cluster</th>
<th>% Future Bad Captured</th>
<th>Cumulative % Captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>11</td>
<td>18621</td>
<td>11625</td>
<td>38%</td>
<td>82%</td>
<td>82%</td>
</tr>
<tr>
<td>12</td>
<td>1388</td>
<td>727</td>
<td>34%</td>
<td>5%</td>
<td>87%</td>
</tr>
<tr>
<td>all clusters</td>
<td>56364</td>
<td>14091</td>
<td>20%</td>
<td>100%</td>
<td>100%</td>
</tr>
</tbody>
</table>

The top two clusters - clusters with highest percentage of future bad customers- are selected for modeling. The cluster with the highest number of future bad customer is cluster 11. This cluster has 30,246 customers. This cluster includes 11625 (38%) future bad customers. Therefore 82% of all future bad customers are in cluster 11. The second worst cluster (cluster 12) includes 2,115 customers including 34% future bad customer. 5% of all future bad customers are in this cluster. Overall these two clusters include 87% of the total future bad customers. There is sufficient information about the future bad customers in these two clusters to justify separate risk models for each of these clusters and clearly there is no problem with data imbalance in either of these clusters.

Each of these clusters is unique. Within each cluster customers have similar transaction and behavior characteristics. As said before modeling each of these dissimilar clusters separately results in discovering additional and unique characteristics of future bad customers.

*Step b) Modeling based on each of the top bad clusters:*
As seen in Table 7.23, cluster 11 includes 82% of the total future bad customers. Customers in this cluster are selected for separate modeling. The dataset is subdivided into training (60%), validation (20%) and test datasets (20%) before model Classification Trees, logistic regression and Neural Networks were applied to this cluster to predict the probability of future bad customers. Table 7.24 shows Gini coefficients for these different techniques.

### Table 7.24  Overall Misclassification Rates for Cluster 11

<table>
<thead>
<tr>
<th>Classification technique</th>
<th>Gini Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>Stepwise logistic regression</td>
<td>0.66</td>
</tr>
<tr>
<td>Classification tree(gini)</td>
<td><strong>0.79</strong></td>
</tr>
<tr>
<td>2 hidden node NN*</td>
<td>0.70</td>
</tr>
</tbody>
</table>

NN*: variable are selected using stepwise logistic regression since there is no inbuilt variable selection option available for NN.

Table 7.24 and the ROC curves in Figure 7.9 show that the Classification tree using gini reduction best identifies risky customers.
Figure 7.9. ROC Curves using different techniques for customers from cluster 11.

The significant variables using the classification tree (gini) model based on cluster 11 are shown in Table 7.25. The whole data set is scored based on this model to predict the Probability of a customer becoming a bad customer within 6 months and the predicted probabilities are denoted as ‘predicted_11’. The same procedure is applied to the customers in clusters 12. The best classifier for cluster 21 was again a classification tree using gini reduction criterion. The significant variables for predicting future bad customers for cluster 12 are also given in Table 7.25.

Table 7.25 The Significant Variables Based on Cluster 11 and 12

<table>
<thead>
<tr>
<th>Significant variables using classification Tree (Gini)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster 11</td>
</tr>
</tbody>
</table>

In cluster 11, customers who spend more than 98% on financial services and have low ratio value for balance at the end of the month to the balance at start of the month are more risky.

In cluster 12, customers with high value of end balance to start balance ratio with high standard deviation are considered as highly risky customers. Also customers who have mean amount of expenditure less than 1 dollar a day and have an AR_7 coefficient (AR coefficient with lag 7) of less than -0.97 are more risky compared to other customers in this cluster.

The models were applied to all customers in the sample data and probabilities of future bad customers based on cluster 11 and 12 trees were predicted for all customers (predicted_11 and predicted_12). The probabilities produced from classifier 11 and 12 along with the other input variables given in Table 6.10 were then inputted to the final model to predict the probability of a future bad customer in September 2007.

**Step c) Final model to predict future performance based on the input variables and the estimated probabilities for the top two clusters:**

Predicted probabilities based on the top two bad clusters were added to the other existing input variables produce 53 input variables for the final modeling of the future risk using the sample dataset with 20% future bad customers.

The sample dataset was partitioned into a training dataset (60%), a validation dataset (20%) and a test dataset (20%). As before using appropriate variable selection methods different techniques namely Neural Networks, Classification Tree and Logistic Regressions were applied to predict the probability of future good/bad customers.
Figure 7.10 shows the ROC curve for stepwise logistic regression, a classification tree using the gini criterion and neural network with 5 hidden nodes.

The above Figure shows that the classification tree performs worse than the other methods. The ROC curve for Logistic regression and neural network are almost overlapping suggesting that the performance of these two techniques is almost the same. The Gini values for these methods are shown in the Table 7.26.
Table 7.26  \textit{Gini Coefficients for Different Techniques}

<table>
<thead>
<tr>
<th>Classification Technique</th>
<th>Gini Coefficient</th>
</tr>
</thead>
<tbody>
<tr>
<td>NN(5hn)</td>
<td>0.820</td>
</tr>
<tr>
<td>Logistic regression(stepwise)</td>
<td>0.816</td>
</tr>
<tr>
<td>Tree(gini)</td>
<td>0.720</td>
</tr>
</tbody>
</table>

The Gini coefficient for classification tree is the lowest amongst the above techniques. The NN with 5 hidden nodes and stepwise logistic regression perform equally well. However, a NN is not easily interpretable and it has a longer execution time. Logistic regression is easily interpretable and well understood. Therefore stepwise logistic regression is selected as the final classification model.

The significant variables for the prediction of future bad customers using the stepwise approach as the final model in the proposed hybrid SOM based classification model are given in Table 7.27.

Table 7.27  \textit{Significant Variables Using Hybrid SOM based classification Model}

<table>
<thead>
<tr>
<th>Variables</th>
<th>Coefficient</th>
<th>P value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intercept</td>
<td>-4.0394</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>AR1_6</td>
<td>-0.8762</td>
<td>0.0001</td>
</tr>
<tr>
<td>AR1_5</td>
<td>-0.6038</td>
<td>0.0252</td>
</tr>
<tr>
<td>HEALTH_BEAUTY</td>
<td>-0.0285</td>
<td>0.0077</td>
</tr>
<tr>
<td>CLOTHING</td>
<td>-0.0215</td>
<td>0.0003</td>
</tr>
<tr>
<td>CONSTRUCTION_MAINTENANCE</td>
<td>-0.0203</td>
<td>0.001</td>
</tr>
<tr>
<td>AUTOMOTIVE</td>
<td>-0.0106</td>
<td>0.0005</td>
</tr>
<tr>
<td>MEDICAL</td>
<td>-0.0089</td>
<td>0.0133</td>
</tr>
<tr>
<td>RETAIL</td>
<td>-0.0077</td>
<td>0.0094</td>
</tr>
<tr>
<td>pay_bal_ratio</td>
<td>0.0069</td>
<td>0.0189</td>
</tr>
<tr>
<td>FINANCIAL_SERVICES</td>
<td>0.0171</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>CASH_ADVANCE</td>
<td>0.0218</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Predicted_12</td>
<td>1.7716</td>
<td>&lt;.0001</td>
</tr>
<tr>
<td>Predicted_11</td>
<td>2.9163</td>
<td>&lt;.0001</td>
</tr>
</tbody>
</table>
Positive coefficients are associated with increased risk and negative coefficients are associated with decreased risk. AR_5 and AR_6 are the autoregressive parameters of lag 5 and lag 6 extracted from the daily transaction time series data for each customer. The significance of these variables highlights the usefulness of model based parameter extraction from time series. However, it is difficult to determine the relationship between the frequency of shopping and financial risk based on these variables. Predicted probabilities of future risk based on cluster 11 and 12 are amongst the significant variables which show the usefulness of cluster based classification. Pay_bal_ratio is the ratio of the payment to the amount of money owed by the customer at the start of the month. HEALTH_BEAUTY describes the percentage of money spent on beauty or health products such as beauty parlors and gymnasium. CLOTHING represents a description code which relates to percentage of money spent on clothing stores, accessories, shoe stores and other related products. CONSTRUCTION_MAINTENANCE represents the percentage of money spent on construction and maintenance such as building materials stores, paint and wallpaper stores and other similar stores. AUTOMOTIVE describes the percentage of money spent on automotive products such as Auto & Home Supply Stores, Motorcycle Dealers and Automobile and Truck Dealers. MEDICAL represents a description code which relates to the percentage of money spent on medicine, doctors and medical procedures. RETAIL is a description code describing the percentage of money spent on retail products such as florist, bookstores and newspaper stores. FINANCIAL SERVICES is one of the created “description codes” indicating the percentage of money spent on government services, government loans, tax preparation services and credit union/credit union/money order services. CASH ADVANCES are the ATM cash withdrawals or cash withdrawn from banks. These extra variables identify the behaviors associated with the behavior of bad customers outside of the top 2 bad clusters. The results in Table 7.27 indicate that the risk is lower for customers who spend more on health and beauty products, clothing, construction, medical and retail stores.

Performance Measure:
The Gini coefficient and confusion matrix for the test data using Hybrid SOM based Classification Model is calculated. Table 7.28 shows the confusion matrix using a cutoff value of 0.5 for the test dataset.
Table 7.28  
Confusion Matrix using Hybrid SOM based classification Model

<table>
<thead>
<tr>
<th>Actual</th>
<th>Predicted</th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Good</td>
<td>Bad</td>
<td>Total</td>
</tr>
<tr>
<td>Good</td>
<td>15829</td>
<td>1181</td>
<td>17,010</td>
</tr>
<tr>
<td></td>
<td>93%</td>
<td>7%</td>
<td></td>
</tr>
<tr>
<td>Bad</td>
<td>1245</td>
<td>2881</td>
<td>4,126</td>
</tr>
<tr>
<td></td>
<td>30%</td>
<td>70%</td>
<td></td>
</tr>
</tbody>
</table>

PCC=86% , cut-off probability=0.5

Using the Hybrid SOM based Classification Model with cutoff value of 0.5, 70% of the bad customers are correctly classified and only 7% of the good customers are incorrectly classified. However if the cut-off value is adjusted according to predicted probabilities the models’ accuracy increases. Table 7.29 shows a range of thresholds.

Table 7.29  
Top Percentiles for the Estimated Probability of a Future Bad Customer

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Future Good</th>
<th>Future Bad</th>
<th>% good classified as bad</th>
<th>% bad captured</th>
</tr>
</thead>
<tbody>
<tr>
<td>1%</td>
<td>3</td>
<td>208</td>
<td>0%</td>
<td>5%</td>
</tr>
<tr>
<td>2%</td>
<td>22</td>
<td>400</td>
<td>0%</td>
<td>10%</td>
</tr>
<tr>
<td>3%</td>
<td>53</td>
<td>581</td>
<td>0%</td>
<td>14%</td>
</tr>
<tr>
<td>4%</td>
<td>106</td>
<td>739</td>
<td>1%</td>
<td>18%</td>
</tr>
<tr>
<td>5%</td>
<td>144</td>
<td>912</td>
<td>1%</td>
<td>22%</td>
</tr>
<tr>
<td>6%</td>
<td>192</td>
<td>1076</td>
<td>1%</td>
<td>26%</td>
</tr>
<tr>
<td>7%</td>
<td>233</td>
<td>1246</td>
<td>1%</td>
<td>30%</td>
</tr>
<tr>
<td>8%</td>
<td>295</td>
<td>1395</td>
<td>2%</td>
<td>34%</td>
</tr>
<tr>
<td>9%</td>
<td>391</td>
<td>1511</td>
<td>2%</td>
<td>37%</td>
</tr>
<tr>
<td>10%</td>
<td>457</td>
<td>1656</td>
<td>3%</td>
<td>40%</td>
</tr>
</tbody>
</table>

Total good =17,010; Total Bad =4,126

In the above Table the top 10 percentiles with highest predicted probabilities are shown with the number of bad and good customers included in each percentile. The test data
has 21,136 customers including 4,126 future bad customers. The top 1% of the customers with highest predicted probabilities includes 208 future bad customers and only 3 future good customers. If we consider the top 1% of the customers as a threshold, 5% of bad customers are correctly classified and only 3 good customers are incorrectly classified as bad. If the threshold is set to the top 5% of the customers with the highest predicted probabilities, then 1,056 customers are considered as future bad customers. Amongst these predicted future bad customers 144 customers are good customer incorrectly classified as bad and 912 future bad customers are predicted correctly. These correctly classified bad customers make up 22% of all bad customers. Therefore we capture 22% of the future bad customers in the top 5% of the customers with highest predicted probabilities. The top 10% of the customers with highest predicted probabilities include 2,113 customers. Considering the top 10% of customers as a threshold we correctly classify 1,656 future bad customers and 457 future good customers are incorrectly classified as bad. Considering the top 10% of the customers with highest predicted probabilities as a threshold we capture 40% of all future bad customers using the proposed Hybrid SOM based Classification Model.

The ROC curve for the Hybrid SOM based classification model is shown in Figure 7.8. The area under the curve can be viewed as the proportion of correct classification across all the possible thresholds.

![ROC Curve](image)

Gini Coefficient = 0.81

*Figure 7.8* ROC Curve for Hybrid SOM based Classification Model using test data.
The Gini coefficient for the training dataset is 0.82 and for the test dataset it is 0.81. The evaluation Table for Gini coefficient in chapter 3(Table 3.3) shows that a Gini coefficient of greater than 0.80 indicates very good classification quality. The proposed Hybrid SOM-based Classification Model is therefore a very good model for behavior scoring to predict the future risk 6 months in advance.

**Comparison of 4 month time horizon and 6 month time horizon**

Looking at the significant variables for the prediction of future risky customers 4 month in advance and 6 month in advance, we can identify many similar variables. The created variables such as clothing, construction and maintenance, medical, health and beauty, cash advance and financial services are common to both the time horizons. Customers who spend more on clothing, medical services, clothing stores, etc are customers who are financially sound and therefore can afford to spend money on such products. Customers removing cash from ATM (cash advance) are more likely to be risky and also customers who use their credit card mainly for products such as government loans, tax payment and other government services are more likely to be financially in trouble and therefore are more likely to default and become future bad customers. These created variables had the same impact on predicting risk 4 month and 6 months in advance.

This indicates that the significant variables do not change and one can rely on them to predict the future risk. However, it should be noted that 6-month predictions are a little less reliable than 4-month predictions. Nevertheless, the respective Gini coefficients of 0.86 and 0.82 are both very good.

**Note**: this model was also performed on dataset including only the good (missed no payments) and slow (missed only one payment) customers. When we only consider customers who are currently good or missed maximum one payment it is not easy to capture the future performance since a very small proportion of these customers are likely to default in the future. This make it harder to predict the performance, however,
our hybrid SOM based classification model had a respectable Gini coefficient of 0.69 in this situation (see appendix G).

7.5 Summary

In this chapter we have developed a model to predict future performance of customers. Future performance is measured based on transaction behavior of customers in March 2007 to predict the risk of customers within 4 months and also 6 months. Customers are then classified as future good or future bad customers based on the predicted probabilities. Higher probabilities indicate higher risk which points to a future bad customer. Created variables in previous chapter were used to predict future risk using a conventional stepwise logistic regression. The performance measures resulted in a good model. We proposed a new Hybrid SOM based Classification Model in this chapter. This model was applied to the same dataset and the performance of the model was compared to that of the conventional stepwise logistic model. The Hybrid SOM based Classification Model has a much better performance and is classified as a very good behaviour model. It is robust to changes in the percentage of bad future customers and to the forecast horizon. In addition it has been shown how a more informative expected loss measure can be used to identify for the bank customers who are most likely to really hurt the bank when their financial situation deteriorates.
CHAPTER 8
CONCLUSION

If you tell people where to go, but not how to get there, you’ll be amazed at the results.

American soldier George S. Patton (1885-1945)
8.1 Introduction

The financial crisis of 2008 resulted in the collapse of large financial institutions, the rescue of banks by national governments, and decline in stock markets around the world. The crisis resulted in failure of key businesses, declines in consumer wealth and a significant decline in economic activity. Experts have suggested several causes and many solutions have been implemented or are under consideration, but still the world economy is under significant risk. Some critics argued that investors and credit bureaus did not accurately price the risk involved with financial products related to mortgages, and that governments failed to adjust their regulatory practices to deal with financial markets during the 21st century. Sound understanding and accurate evaluation of the risk is therefore a critical part of financial institutions. This is one of the key factors that motivated this work.

We believe that if a credit card customers’ financial situation is changed, then the customers’ transaction behaviour changes. If these behavioural changes are captured quickly then the bank is timeously warned about changes in their risk exposure and can make appropriate changes in terms of customer management in order to avert any future problems.

Credit scoring is a discipline devoted to developing statistical models that guide organizations to make appropriate credit decisions. Credit scoring uses predictive modeling techniques to analyse historical data and transform this historical behaviour into numerical measures that quantify the likelihood of future performance. For example, based on transaction and payments behaviour, history of non-payments in the past 12 months, credit utilization and other factors, customers are scored to predict the likelihood of customers defaulting on their credit facilities in future.

Credit scoring has become an essential process in the financial industry. Credit scoring has many benefits including closer monitoring of existing accounts, reductions in credit analysis costs, quicker credit decisions and prioritizing credit collections (Brill, 1998).
It is a common practice in banks to apply statistical analysis to classify existing customers into future good (not-risky) and future bad (risky) customers. However, this is done using demographic information, historical delinquency status (number of times payments are missed in the past year), credit utilization and other credit scores such as application scores etc. Historical delinquency status, utilization, demographic information and other credit scores are highly predictive of future performance. However, to get this information, banks have to extract data from different data sources. This task is not easy and sometimes it takes a long time before all the required data is available. Moreover, there are many issues with the data including many missing values. According to many practitioners data preparation and cleaning is very time consuming and is a very crucial step before model development.

Hand and Henley (1997) have most appropriately advised that in order to build a more accurate credit scoring model concentration should focus on new and more predictive variables rather than the method of analysis. In this research we have concentrated on creating new input variables based on easily available transaction and payment history which are predictive of future performance. Then we use a new approach to create other new variables which are combinations of other important risk predictive variables. These new combined variables along with other variables are used to predict the probability of future bad customer. Results show that these new combined variables improve the predictive accuracy of the model.

The financial savings and opportunities from high-performing scorecards are very important in today’s world. With the growth in the number of credit cards even slight improvements in the model accuracy can increase profit / decrease losses appreciably.

The entire development process of a practical scorecard was examined in this study using a large set of real-world credit scoring data from a leading Australian bank. The goal was to build a credit scoring system that is different to the systems presently used at the bank. The major processes involved in building a credit scoring system (data preparation and cleaning, variable selection, samples generation, model development and validation) are explained in this thesis.
Performance of different statistical and data mining classification models viz., logistic regression, decision tree and neural network are compared and a new hybrid model is used to develop a high-performing scorecard.

Our model has proven to be an accurate model in classifying future good and bad customers and should therefore result in financial savings and better strategy planning for banks. Several financial institutions have already displayed significant interest in the work that has been done.

8.2 Thesis Summary

Behaviour scoring models aim to group customers that share similar behaviour patterns. Using these patterns, banks target different groups to promote new products and increase credit limits, encouraging these groups to spend more. In addition banks target financially unhealthy groups, coming up with strategies to manage recovery if a customer’s repayment ability turns bad. In behaviour scoring models, historical transaction behaviour and payments are considered assuming that the customer’s behaviour will be similar in the future. In order to model a customer’s behaviour pattern, behaviour scoring models establish an association between input variables and an output score, which measures the probability of risk. Based on these associations, a score is assigned to each customer and customers are clustered into groups for management purposes.

In this research, the risk behaviours of credit card customers were modeled at an individual level to predict longer-term changes in behaviour before any activity damaging to the bank has occurred. The intention is to enable the bank to conduct preemptive action minimizing monetary losses. The chapter summaries are as follows:

Chapter 1: Introduction

- Banking culture is introduced to the reader emphasizing the impact of credit cards on the banking industry. The aim of the project is clarified and an overview of the thesis is discussed. The three main aims of the project were highlighted in this chapter. These aims were to show:
The usefulness of time series clustering based on a mixture of model-based and feature extraction methods in predicting customers’ future performance in behaviour scoring systems.

The usefulness of product identified expenditure in predicting customers’ future performance in behaviour scoring systems.

The superiority of a proposed Hybrid SOM based Classification Model as a behaviour scoring model over conventional (logistic regression) classification techniques.

Chapter 2: Literature review
- The available literature on credit scoring is discussed in chapter 2. Starting with an overview of credit scoring, we discuss the effect of credit cards on consumer credit. Behaviour scoring theory is explained in detail and past scoring systems are reviewed. Credit scoring literature is discussed in relation to seven commonly used approaches for predicting delinquency namely, linear discriminant analysis, regression, logistic regression, decision trees, non-parametric models, genetic algorithm and artificial neural networks. Also exploratory approaches for credit scoring data are discussed. Many comparison studies are reviewed and advantages and disadvantages of the common techniques are talked about.

Chapter 3: Developing a credit scoring system
- The detailed process of steps involved in developing a credit scoring system is not readily available in the literature. In this research, the major processes involved in any credit scoring system (data preparation and cleaning, variable selection, samples generation, model development and validation) were explained in detail to gain a better insight into the world of credit scoring modeling. First we discuss the data preparation and data cleaning which is one of the most important and highly time consuming steps. Common variable selection methods are discussed and the problems with sample generation are
talked about. After talking about model development and validation, we discuss the two most common performance measures; the percentage correctly classified (PCC) which is very common in the academic literature and the ROC Curve and Gini coefficient which are the main performance measures in industry. We refer to authors who have exposed the flaws of PCC in credit scoring and we demonstrate this point while analysing the main dataset.

Chapter 4: Methodology

- Techniques used for both the studies in this thesis are explained in chapter 4. Time series clustering in explained briefly highlighting the advantages of time series clustering based on a mixture of model based characteristics and feature extraction. Two different clustering techniques are used in this thesis. In the first study we have a small dataset and therefore we have used Ward’s hierarchical clustering method. In the second study the dataset is very large. Self Organising Maps are used to cluster customers mainly because of their visual advantages. Both these clustering methods are explained in chapter 4. Autoregressive models which are used to describe our time series data are explained before we move to classification methods. LDA, Logistic Regression, Classification Trees, Neural Networks and two stage hybrid models are explained in detail. Also, the advantages and disadvantages and suitability of the techniques for credit scoring are discussed in this chapter.

Chapter 5: Pilot study

- A small banking dataset was provided to us with transaction amounts for 1000 customers, over 6 months. The aim of this pilot study was to gain more knowledge about customer behaviour for marketing purpose. Customers were segmented into different clusters recognizing big spenders and small spenders. A new approach was applied to prepare new predictive variables for modeling. In this analysis the clustering of customer time series was based on a mixture of the model approach and feature extraction approach for each time series. Customers’ transactional data were monitored to expose the changes in spending patterns over time based on the type of products purchased. This cluster analysis was repeated each month using hierarchical clustering. The
movements of customers between the clusters were monitored to gain an understanding of the changes in customers’ behaviour over time.

Chapter 6: Data preparation for the main study

In chapter 6 we create the variables for the main study. We believe customers who are financially in a good position behave differently to customers who are financially in trouble. Credit card customers default mainly because their financial circumstances changes. These changes result in unusual transaction behavior, which allows the identification of financially troubled customers. New predictive variables are created based on type of products, the daily transaction amount and the repayment amounts. The daily transaction amounts are considered as time series. We aimed to group similar behaved customers based on the daily transaction amount, percentage spent on different products and repayment habits. However, the time series are not directly used for clustering. A model is fitted to each of the series and also some features are extracted from these daily time series, allowing us to group the customers on the basis of their financial behaviour. Based on model parameters fitted to these series, features extracted, percentage spent on different products and repayment ratios we aim to group customer and then predict the future performance of these customers. All the created variables have an information value greater than 0.1 in the SOM which indicates that all these variables are useful for the prediction of future performance. The variables used in this thesis are very different to the variables used in conventional credit scoring. One of the longest steps in building a scorecard is the data extraction step. We have used information which is readily available so minimal data extraction effort is required. This makes the building of a credit scoring model much quicker with our new set of variables.

Chapter 7: Main study

For the main study, we were provided with a large dataset from a leading Australian bank. This data included the transactional history of 2.4 million customers. The aim of this study was to predict the probability of customers turning delinquent in the future. Future performance of customers is predicted
based on the created input variables and customers are categorized as risky (future bad) or non-risky (future good) customers. Two different models are applied to predict the future performance: A conventional stepwise logistic regression model is applied to demonstrate the significance of the created input variables in predicting future bad customers four months ahead. Then a hybrid SOM based classifier is proposed to predict the probability of future bad customers and the amount owed by these customers. The superiority of the proposed classifier is verified by comparing the performance of the two models. In this study the significance of Autoregressive parameters and moments extracted from the transaction time series show the usefulness of time series clustering based on a mixture of model-based and feature extraction methods in predicting customers’ future performance in behaviour scoring systems. Also the high prediction accuracy of the model demonstrates the usefulness of product based expenditure in predicting customers’ future performance in behaviour scoring systems.

8.3 Contribution:

When the research aim was discussed with the above Australian bank, the bank was very interested and mentioned that, due to time constraints, they were not able to apply novel sophisticated approaches for building credit scoring system.

Useful practical information was obtained during discussion with credit risk modeling practitioners and a large set of real world credit scoring data was provided to us for modeling with this purpose in mind. In particular it was stressed that this research needed to be relevant and easy to implement within existing scorecard development processes.

Unlike other credit scoring systems, our credit scoring system is not based on any socio-demographic variables. In this study, we have only used the transactional data including the amount of money spent on each transaction, the payment amount, delinquency status and the merchant code where each transaction took place. No demographic information
or information about credit limit or credit card type was provided. Despite this limited range of input data our model easily matches the performance level of other models in terms if the accuracy of its delinquency predictions.

Despite widespread literature on credit scoring, most of the studies deal with benchmark datasets. Few studies are conducted on real-world datasets and there is a lack of studies in the literature that explain the entire process of scorecard development, from the raw data processing to the final model. There is a need to develop new approaches, with new variables which do not rely on the arduous binning of data and univariate credit scoring approaches currently used by banks. These approaches do not utilize modern statistical developments and data extraction capabilities as much as they should. However, although our study uses multivariate methods that do not require binning the concerns of practitioners were considered. Discussion with practitioners and focusing on the main purpose of building a credit scoring system was very important to ensure the successful implementation of our new approaches.

For example discussion with practitioners suggested that it was not appropriate to use a classification tree as a final model for predicting delinquent customers because equal probabilities would be assigned to all the customers in each child node. Tree classification has been successfully used in building many credit scoring systems and has been selected as the best model when they had the lowest misclassification rate compared to other models (Srinivasan & Kim, 1987). However, the delinquency probabilities obtained using classification trees are equal for all the members of same child node so that we cannot differentiate between customers in the same child node. This is the reason why classification trees are not used in practice despite the fact that classification trees permit the modeling of complex dependencies between variables (Srinivasan & Kim).

We therefore present a new credit scoring system that uses a mixture of techniques: First clustering of time series is performed to group similarly behaved customers, then a classification tree is applied to the clusters with the highest percentage of bad customers. The outcome of these models is the estimated probability of each customer becoming delinquent in the future (4 months or 6 months) based on selected highly predictive variables. The predicted probabilities based on each of these clusters are
considered with other input variables for the final logistic regression model. A final
customer risk evaluation is obtained from a two stage model, predicting the probability
of default at the first stage and the amount owed by the customers at the second stage.
This hybrid model has proved to be very accurate and successful for evaluating future
risk. Interaction effects were taken care of in the models for the high delinquency
cluster, thereby allowing for a relatively simple logistic regression final model.

Behavioural scoring helps banks identify changing characteristics of existing borrowers’
behaviour (Thomas, 2000; Thomas, et al., 2001). Hopper and Lewis (1992), Hsieh
(2004) and McNab and Wynn (2000) give more careful accounts of how behavioural
scoring systems are used in practice.

Developing an accurate credit scoring model helps banks and other credit institutions to
protect themselves from customers with high default risk. However, as pointed out by
Lim and Sohn (Lim & Sohn, 2007), current available scoring models have some
shortcomings that need to be addressed.

First of all, the current scoring models require a certain length of observation period to
predict the customers’ future risk performance. The lenders have to wait until the
information for the full observation period is collected. However, usually lenders want
to know if the borrower is risky as soon as possible and it is not possible to wait (Lim &
Sohn, 2007).

In this research we have proposed a new approach which observes behaviour for
only a one month period before making a prediction for the risk of delinquency.
Therefore the lenders do not need to wait to collect the information for a longer
observation period. The transactional data is monitored in such a way that the
financial circumstances of a customer can be measured for the present month. The
financial behaviour of the customers for the current month is then used to predict
their financial behaviour in the future.
Current available models apply a single classification rule to every customer in the same way. A single classification rule cannot recognize the fine distinctions between the characteristics of different types of individual borrowers. Therefore this kind of single rule classification model can be inefficient (Lim & Sohn, 2007).

In this research we have proposed a new approach for variable selection and a classification rule that can identify more of the significant characteristics of the bad (risky) customers. We have grouped the customers into several clusters. Each cluster is unique and has its own distinctive characteristics allowing us to identify the most risky clusters. Applying a different classification rule to each of these clusters separately helps identify the characteristics of bad customers within each cluster. Combining all these classification rules we gain better insight on the characteristics of the bad customers producing more accurate predictions of delinquency. This point is explained in detail in section 8.3.3.

It has been found in the literature that the use of new and more predictive variables can improve the performance of scorecards (Hand & Henley, 1997). However it is difficult to obtain large real world credit scoring datasets containing many variables. This is the reason why only a few studies have examined the use of variable selection techniques in credit scoring.

In this research we have used a very large real world credit scoring dataset from a leading bank of Australia containing a limited selection of variables. We have created new variables which we have found to be useful in identifying the financial circumstances of existing customers. This point will be explained in detail in section 8.3.1.
Banking credit scoring systems are commonly used for large portfolios. This means that a slight improvement in credit scoring accuracy can reduce the lender’s risk and translate significantly into future savings. Developing an accurate credit scoring model helps banks and other credit institutions to protect themselves from customers with high default risk.

In this research we have developed a new hybrid model, which combines several techniques to increase the accuracy of the model. A two-stage model is then used as the final model to predict the delinquency probability at first stage and balance owed at second stage. The usefulness of this method is explained in detail in section 8.3.2.

In this research we have contributed to the theoretical and practical aspects of credit scoring systems. Theoretically we have developed a new approach in data preparation, variable selection and modeling. Practically we have developed a credit scoring system that is completely different to the existing credit scoring systems and we have developed the SAS code to implement this system. This approach is of great interest and the bank that supplied the data is looking forward to finding ways to implement this approach.

The information used in this research could be considered as ‘positive credit reporting’, a new concept in Australia. Positive credit reporting allows credit providers and firms that provide services on credit to share information on individual consumers' credit files by subscribing and contributing this information to a central database. Shared information may include amounts borrowed, type of credit facility, credit provider, repayment history on all credit accounts and default information. The Australian Government is reviewing the current process of the Privacy Act to consider amendments which will facilitate the adoption of 'Positive Credit Reporting' structures. Currently Positive Credit Reporting is used in the US. This new structure allows credit reports to include a much broader range of information on outstanding balance, credit limit and payment history. No negative information such as number of times payments are missed in the last one year is used in positive credit reporting.
Access to positive credit reporting has many benefits for the financial institutions. Along with many other benefits, positive credit reporting can improve risk assessment for lenders, enable pricing to more accurately reflect risk and also prevent over-commitment, bad debt and fraud. This means that in the future exploiting the type of data used in this study will not be limited to banks. As a result this work is likely to have much wider areas of application in the future.

The above contributions introduced in this section are explained in detail in the following section.
8.3.1 Data preparation

The ‘Data preparation’ step is a very important step in credit scoring. It has been found in the literature that the use of new and more predictive variables can improve the performance of scorecards (Hand & Henley, 1997). One of the main contributions of this thesis is the creation of new input variables for modeling. These variables have proven to be highly significant for the prediction of delinquency.

c) Merchant Category Code

We have created new variables which we believe have the potential to explain the financial circumstances of customers. The amount spent every month is subdivided into categories. These categories are selected in such a way that we can differentiate between financially secure customers and customers who are not financially healthy. We believe that if a customer is not financially sound, he will not spend money on luxury products such as sky diving and jewelry; instead his credit card will be used for more necessary products such as petrol and food. The amounts spent on different types of product are considered and the products are grouped in such a way that we can identify the financial circumstances of customers.

In the first analysis percentages spent for different product categories are calculated every month and changes in those percentages are monitored each month. This helps banks to identify changes in financial behaviour every month. In the second analysis, amounts spent for each product category provide input variables for clustering and for developing models for delinquency risk. Based on where and how much the customers spend, similar customers are grouped and then scored in terms of delinquency risk.

d) Autoregressive parameters and moments:

Contrary to common practice, transaction amounts are not directly used for clustering, in this study. Instead a mixture of model based and feature based approaches are applied to daily transactional time series data in order to summarize this data. Due to the weekly patterns observed in many series, AR (7)
models are fitted to the daily time series information for each customer, resulting in 7 parameters. In addition, time series features such as the mean, standard deviation, skewness and kurtosis are extracted for each customer. The extracted features and estimated parameters (along with percentage expenditure per MCC) are used as input variables for clustering. Few of these Autoregressive parameters and moments prove to be significant in prediction of delinquency; however, the delinquency rates for the resulting clusters differ dramatically. An additional advantage of the proposed approach for time series clustering is that this approach is appropriate even if we must deal with time series of different lengths. Most institutions have separate score card systems for old customers who have been with the bank for more than one year and customers who are new at the bank since the transactional data for old and new customers differ in length. Using a mixture of model based and feature based time series clustering, a model can be developed for both new customers and old customers simultaneously, since parameters could be estimated regardless of series length. However, our models require only one month of data, making them immediately applicable to old and new customers.

8.3.2 Model Development and Validation

The ‘model development and validation’ step is used to discriminate between ‘good’ and ‘bad’ customers. The better the classifier, the better will be the performance of the scorecard. To select the best model for the underlying data, different models should be applied and compared to identify the best performing model. Using SAS Enterprise Miner, we can easily compare a variety of model types such as decision trees, logistic regression or neural networks. In SAS Enterprise Miner predictive performance can be assessed and compared for these models using a holdout or test dataset.

It has been found in the literature that a combination of techniques can improve predictive accuracy (Lee & Jung, 1999/2000). However, there are very few studies which have used a combination of models /mixture of techniques for credit scoring
(Bahrammirzaee, Rajabzadeh, Ahmadi, & Madani, 2011; Chuang & Huang, 2011; Gopalakrishnan, et al., 1995; Lee & Chen, 2005; Lee, et al., 2002). We have used mixtures of techniques and have found that a combination of techniques does result in an improvement in the prediction accuracy of the final model. The final Hybrid model is selected keeping in mind the objective of the research, the identification of risky customers using quick, practical and reliable methods.

In this research we have not only predicted the probability of delinquency, but also we have predicted the balance owed by customers to the bank. We have used a two-stage model for this purpose to first predict the delinquency probability and then predict the balance owed. This can help banks and other financial institutions to identify and then concentrate on rehabilitating the most critical delinquent customers. In other words, the bank can differentiate between bad customers and very bad customers. For example, a customer who does not repay the minimum expected amount for three months is considered as a delinquent customer. However, this amount could be $100, $1000 or $10000. By just predicting the delinquency status one cannot differentiate between delinquent customers who owe a large amount to the bank and those who owe a small amount. Therefore predicting the balance owed by customers is very useful for banks. Having more knowledge about their risk exposure and a better understanding of the behaviour of more risky customers allows banks to come up with strategies to reduce future losses (bad debts).

The validation of the hybrid SOM based models developed has involved

a) The entire dataset using all the customers who had transactions in March 2007. Excellent classification results were obtained.

b) The entire dataset using all the customers who made a transaction in April 2007 and also May 2007. This is called an out of sample validation. Out of sample indicates that the customers in this dataset were not included in the model development dataset. Excellent classification results were obtained.
c) Both 4-month and 6-month forecast horizons, producing excellent classification results in both cases.

The above models were developed using a random sample of March 2007 customers. In separate analysis the model was applied to the customers who had missed at most one payment by March 2007 and the performance of these customers were predicted in future. When we only consider customers who are currently good (missed no payments) or missed maximum one payment (slow payers) it is very hard to capture the future performance since a very small proportion of these customers are likely to default in the future. This make it harder to predict the performance, however, our hybrid SOM based classification model had a respectable Gini coefficient of 0.71 in this situation.

The good results obtained for all four of these validation exercises confirms that the hybrid SOM based classifier is a robust tool which can be relied on in many different contexts.

8.3.3 Variable Selection approach

Most credit scoring studies use benchmark datasets which are already clean and contain only a few variables. This is because it is difficult to obtain large real world credit scoring datasets containing many variables: Benchmark dataset studies do not require variable selection methods because there are so few variables. Only a few studies have therefore examined the use of variable selection techniques in credit scoring.

A new approach for variable selection is developed in this study. We segment the customers based on their behaviour in only one month. Each cluster is unique and has its own distinctive characteristic. The three clusters with the maximum percentage of delinquent customers within 4 months were selected and for each of these clusters delinquency is modeled separately. This is done to identify the characteristics of bad customers within each cluster. Customers in a particular cluster have similar behavior. Modeling based on a particular cluster can therefore identify characteristics which can differentiate between good and bad customers with similar behaviour. This approach
can help identify the significant customer characteristics for predicting delinquency for various risky concentrations of customers. This information does not emerge to the same extent if modeling is performed on the whole dataset instead of clusters. As said by Lim & Sohn (2007), a single classification rule cannot recognize the fine distinctions of the characteristics of various individual borrowers and therefore this kind of multi-cluster classification model can be more efficient.

In order to identify risky customers we have applied different classification rules (tree, NN, logistic regression models) to each of these clusters separately and the model with minimum misclassification rate was selected in each case. This allows for different levels of interaction/non linearity for the delinquency rules within each cluster. Applying different classification rules to each of these clusters separately helps identify the characteristics of bad customers within each cluster. Models based on each of the three clusters are applied to the whole dataset and predicted values (predicted_n; n =1, 2, 3) along with other variables are used as input for the final classification model developed for all customers. The use of this hybrid model helps in identifying the different characteristics of similarly behaved good and bad risk customers and also the different characteristics of good and bad risk customers outside of the 3 riskiest clusters.

**8.3.4 One Month Observation Period**

The key assumption of credit scoring models is that the future is like the past (Berry & Linoff, 2000). This assumption is not always true. There can be several time lags between the collection of transactional data and its use in a scorecard. A scorecard used in 2010 is likely to be built in 2009 and in order to have enough history in a sample to decide if the customers are good or bad, a sample of customers’ behaviours from 2007 and 2008 would be considered. During this time, many changes may have occurred in the financial situation of customers. This is known as population drift. The impact of population drift is of interest to both academics and practitioners (Hand & Henley, 1997).
Population drift occurs because populations change over time. These changes could be due to changing economic conditions or changing customer perceptions on borrowing (Thomas, et al., 2004). These changes can cause changes in the distribution of population characteristics and hence a model’s prediction capability will usually diminish over time. Hence credit scoring systems need to be adjusted frequently to consider new situations.

The proposed approach avoids population drift. In this study behaviour of customers is based on what products they buy and how they pay for these purchases in only the last month. The analysis considers only one month of transaction data to predict behaviour 4 (and 6) months later. Therefore the data used in building this scorecard is very recent data which identifies the present circumstances of the customer. Our models suggest that financially sound customers do not spend on luxury products such as gymnasium, jewels, water sports etc. In fact a customer who is financially in trouble uses his credit card carefully. Cash advances and financial services tend to feature prominently representing high risk. On the other hand expenditure on medical, education, automotive is indicative of low risk.

Findings of this research are summarized below;

- A short(1 month) development sample is used to develop the scoring model
- New variables are created for modelling
- A new method (hybrid SOM) is used for classification

Overall we have come up with a new scoring system which is based on a customer’s consumption type and very recent transactional history to understand the current financial circumstances of customers.
8.4 Limitations of the Research

The results show that the proposed Hybrid SOM based method is successful for credit scoring; while time series clustering has been successful in finding new variables for identifying delinquent customers. However, there are a few limitations in this research.

We did not have access to the credit limit of customer credit cards. Knowing the credit limit we could have created new variables to indicate what percentage of their credit limit customers are using. The addition of this variable may have been useful for identifying delinquency, thereby increasing the predictive accuracy of the model.

We also excluded fines and fees in this research. This was perhaps not a good idea since customers who are financially in trouble change their repayment habits and start missing payments. These are customers who get fined and therefore including fines could be an advantage in identifying delinquent customers. This would be too late for intervention however. We want to predict delinquency before it strikes.

Also the created variables could be broken down to smaller groups which may be more representative of risk. For example, AUTOMOTIVE represents the money spent on car purchases and also car repairs. Separating the car purchases from the car repairs may be very useful in identifying risky groups.

The type of customers considered in this research was limited to credit card customers and data for all other banking services was ignored. Clearly it would have been better if other bank information could have been utilized such as mortgage debt.

8.5 Suggestions for Future Research

The conclusions and findings of this research suggest many challenging issues for further work related to credit risk.

As said before there are two types of credit evaluation (refer to Chapter 1).
One is *behavioural scoring* which has been the main topic of this research. It deals with the supervision of existing customers, including whether or not to increase their credit limits. In this research our primary aim was to predict which of the existing customers are going to default in the future. We were not provided with any information regarding the credit limit of credit cards. This work can be extended to consider increases in credit limit or the offer of new products to existing customers based on the evaluation of risk for each customer.

The proposed ideas and classification model can also be useful in direct marketing to determine which marketing channel (mail, email or telephone) to use for future as well as existing customers. For example it might be found that customers who respond to mailed marketing are more likely to default suggesting that this marketing channel should be avoided.

In this research we did not have access to customer socio-demographic data. Therefore, an opportunity for further research involves the monitoring of customer financial behaviour as well as their socio-demographic status and incorporating both types of variables when modeling future risk.

The other type of credit evaluation is *application scoring* which deals with approval of credit for new applicants. The proposed ideas and method can be used in building an application scoring model. This can be achieved by capturing some of the application scoring variables and building a model based on these variables using the proposed method.

In this competitive world, the objective of credit scoring is now changing from trying to minimize the probability of default to trying to maximize the profit that can be generated from granting applicants credit. As such, another avenue for future work is to investigate the proposed model in the context of *profit scoring*.

It would be particularly interesting to explore how the ideas that were applied in a behaviour credit scoring setting could be translated to application and profit scoring settings.
As explained in chapter 2 Genetic algorithms (GA) are very useful in classification problems. South et al., (1993) highlights the reasons why GAs are useful in classification problems such as credit scoring (refer to section 2.5), also the effectiveness of GAs in searching large datasets is discussed by Goldberg (1989). GAs learn complex relationships in incomplete datasets and can discover unknown patterns (Goonatilake & Treleaven, 1995) and hence they are highly suitable for credit scoring. GAs has been successfully used as scorecards in a credit union environment (Desai, et al., 1997). However, due to software and time restrictions we could not apply GA to the dataset in this research.

It would be particularly interesting to explore the performance of GA and to compare it with the other methods discussed in this research.

Another interesting future research area only touched on in this research is the movement of customers between clusters from one month to another, perhaps using Markov Chains.

In this research we have looked at one month behavior of customers to predict the future risk 4 months and 6 months in advance. However, considering only one month we may not capture all the spending patterns of future good/bad customers. For example, in the one month development sample there were probably not enough customers who were traveling because March 2007 was not in a holiday period. If one has enough customers in every category then it may be possible for more significant variables for risk prediction to be selected allowing for more accurate models to be developed. Certainly more analysis is required in monitoring the performance of models at other times of the year, especially during holiday periods when financial budgets are often stressed.

Another important suggestion for future research is a comparison of length for development samples. Currently banks use 18 to 24 months transactional history to build models. In this research we have pointed out the suggestion of many authors to take a shorter period to overcome the problem of population drift. We have only used one month of transactional behaviour for our development sample and have obtained
good results. Future studies could compare the performance of models using
development samples of different length (for example 1, 6, 18 and 24 months).

8.6 Extensions to Other Contexts

The ideas presented in this research can also be applied to a large number of other areas. One such example is sport. Time series clustering has been used to automatically compile video and audio highlight packages, by gauging audience reactions (Radhakrishnan, Divakaran, & Xiong, 2004; Radhakrishnan, Otsuka, Xiong, & Divakaran, 2005). We have applied some of the ideas and methods used in this research to cricket data to predict the performance of players (Bracewell, Farhadieh, Jowett, Forbes, & Meyer, 2009).

We used time series clustering to characterize Bradman’s cricket career and to compare him with other great cricketers. This paper has introduced a methodology for comparing sporting careers that was designed for the financial sector and has demonstrated that time series clustering is an effective technique for comparing sporting careers. The methodology described in this paper has wider applications, such as comparing players for scouting and selection purposes in a variety of sports.

But there are of course many other areas that could benefit from these approaches. As explained by many researchers, time series clustering is an active area of research, especially in financial and medical applications (Basalto & De Carlo, 2006; Pattarin, Paterlini, & Minerva, 2004; Savvides, Promponas, & Fokianos, 2008). Detection of insurance fraud based on payment and claim history is an obvious example, but the prediction of tax fraud and churn are others likely candidates. Changes in payment and demand patterns may be useful in predicting when people are likely to change their electricity or mobile phone service providers. Any business that collects data from customers on a regular basis should consider extracting time series parameters for the purpose of better understanding their customers.
How this information is used is of course dependent on the type of business. Such information allows the segmentation of customers into similar groups, permitting more focused operational and marketing strategies. This information also allows the prediction of future customer activity and behaviour, giving business an opportunity to control customer activity to a greater extent.

The internet provides a particular opportunity in that customers’ behaviours (clicks etc.) are recorded as time series. The time people spend on various web sites is already used by many customers to promote sales in other areas.

8.7 Summary

Despite widespread literature on credit scoring, as mentioned in the first and second chapters, few studies consider behavioural scoring. Most scoring studies use benchmark datasets which are already clean and contain few variables. Few studies are conducted on real-world datasets and these studies contain only the most relevant variables and therefore there is no need of data preparation or variable selection. Only a few studies have examined the use of variable selection techniques in behaviour scoring. There is a lack of studies in the literature that explain the entire process of scorecard development from the raw data to the final model. There is a need to fill this gap in order to differentiate between the theoretical and practical aspects of scorecard development and to ensure the successful implementation of new approaches.

The typical practical scorecard usually involves the following steps: 1) Data preparation, 2) Data cleaning, 3) Variable selection, 4) Samples generation, 5) Model development and validation, 6) Model approval. All these steps are essential in developing a scoring method. In this thesis all these steps are explained, however, the focus is on the ‘data preparation’, ‘variable selection’ and ‘model development and validation’ steps.

The contributions in this research relate mainly to “data preparation”. We have created new input variables for detecting future risk performance. Also we have contributed to ‘variable selection’ by taking a different approach for the selection of bad customer
characteristics. We use unsupervised SOM to group customers and model three groups with the highest percentage of future bad customers identifying the characteristics of similarly behaved good and bad customers within these groups. We have also contributed to the ‘model development and validation’ by proposing a Hybrid SOM based classifier which has been proven to be a very good behaviour scoring model.

The data for this research came from two important Australasian banks. There were several discussions with bank analysts to ensure that this research was relevant and easy to implement within existing “scorecard” development processes.

Deterioration in the financial circumstances of customers is often associated with default. If this change in financial circumstances can be detected early, banks are able to ensure that they have sufficient capital to cover this risk. The main purpose of this thesis is to identify changes in the financial circumstances of existing customers as early as possible. In this project risk behaviour of credit card customers was modeled at an individual level to predict longer-term changes in behaviour before any activity damaging to the bank has occurred. The intention is to enable the bank to conduct preemptive action minimizing both reputational and monetary losses.

The first study (pilot study) was performed on a small dataset and was limited to only transaction amounts and type of transaction. This pilot research study was useful in getting a better understanding of transaction behaviour. In this study customers are grouped (clustered) based on transaction amounts and the type of product spend for a month. A more sophisticated method of clustering based on extracted global characteristics from time series is applied to the data to extract useful information. This process is repeated separately for 6 months and customers are clustered in each of these months. The movements of customers between the clusters are monitored to gain an understanding of the changes in customers’ behaviour over time.

The purpose of the main study was to identify customers who are likely to miss payments in the future based on their transaction behavior. In the main research study a large datasets is analyzed. New predictive variables are created based on the methods used in the pilot study. Future performance of customers is predicted based on the created input variables and customers are categorized as risky (future bad) or non-risky
(future good) customers. Two different models are applied to predict the future performance. First a conventional stepwise logistic regression model is applied to demonstrate the significance of the created input variables in predicting future bad customers four months ahead. Then a hybrid SOM based classifier is proposed to predict the probability of future bad customers and the amount owed by these customers. The superiority of the proposed classifier is verified by comparing the performance of the two models. In this study the usefulness of time series clustering based on a mixture of model-based and feature extraction methods in predicting customers’ future performance in behaviour scoring systems is shown. The good performance of both the models highlights the usefulness of product identified expenditure in predicting customers’ future performance in behaviour scoring systems.

Current systems are based on a 12 month view of historical customer behaviour and assume that a customer’s future behaviour is similar to past behaviour. In this thesis we develop an early warning system with a historical one month view that fully utilizes all available transactional information collected in the past month and predicts the future performance of customers based on this current information. This time frame can give us a better understanding of a customer’s present financial status and banks can make a quick decision. This one month view is also useful in monitoring the new customers. Current bank systems can only consider customers who have at least a 6 month history. Therefore new customers cannot be scored for the first 6 months. With our model, new and old customers can be modeled at the same time because we require only one month of transactional data for each customer. Therefore with minimal information we are able to predict the performance of all customers in the future.

We have successfully come up with a sophisticated and novel approach not only to satisfy the academic requirements of new research but also to produce better results for the banking industry. On a final note, the critical contributions in this research relate to the discovery of new variables for detecting delinquency, new procedures for predicting future delinquency, and the use of only current data for this purpose.
I almost wish I hadn't gone down that rabbit-hole — and yet — and yet — it's rather curious, you know, this sort of life! I do wonder what can have happened to me! When I used to read fairy-tales, I fancied that kind of thing never happened, and now here I am in the middle of one! There ought to be a book written about me, that there ought! And when I grow up, I'll write one.'

— Alice's Adventures in Wonderland, Lewis Carroll

Thank You for Reading.
REFERENCES


236


Marquez, J. (2008). *An introduction to credit scoring for small and medium size enterprises.*


239


Appendix A

SAS code for extraction of Balance Amount

/********************************************
* Get current months balance for payment/balance ratios
********************************************/
libname raw 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Raw_SAS_Data';
libname out 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Global_characteristics';
libname mccs 'C:\Documents and Settings\Administrator\Desktop\XXX_production\mcc_categories';
libname bal_del 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Balance_delinquency_data';

%let year= 06 07 08;
%let months1=10 11 12;
%let months2=01 02 03 04 05 06 07 08 09 10 11 12;
%let months3=01 02 03 04 05 06 07 08 09;

%macro mnthly_bal_data;
  %let k=25;
  %let kminus4=%eval(&k-4);
  %let kminus6=%eval(&k-6);
  %let i=1;
  %let y=%scan(&year,&i);
  %do %until ("&y"="");
    %let j=1;
    %let m=%scan(&&months&i,&j.);
    %do %until ("&m"="");
      %put k &k. kminus4 &kminus4 kminus6 &kminus6;
      data &bal_del..b&y&m;
      set &input_lib..mis_rolling_bal_anon(keep= ac_key STMTDTE_&k. BALBT_&k. BALCASH_&k. BALRTL_&k. STMTBAL_&k.);
      rename STMTDTE_&k.:=STMTDTE BALBT_&k.:=BALBT BALCASH_&k.:=BALCASH BALRTL_&k.:=BALRTL STMTBAL_&k.:=STMTBAL ac_key=ACCT_NUM;
      run;
    
    data &bal_del..future_b&y&m;
    set &input_lib..mis_rolling_bal_anon(keep= ac_key STMTDTE_&k. STMTBAL_&k. STMTDTE_&kminus4. STMTBAL_&kminus4. STMTDTE_&kminus6. STMTBAL_&kminus6.);
    rename STMTDTE_&k.:=STMTDTE_NOW STMTDTE_&kminus4.:=STMTDTE_4m STMTDTE_&kminus6.:=STMTDTE_6m STMTBAL_&k.:=BAL_now STMTBAL_&kminus4.:=BAL_4_month STMTBAL_&kminus6.:=BAL_6_month ac_key=ACCT_NUM;
  %end;
%mend mnthly_bal_data;

%mnthly_bal_data;
run;

%let k=%eval(&k-1);
%let kminus4=%eval(&kminus4-1);
%let kminus6=%eval(&kminus6-1);

    %let j=%eval(&j+1);
    %let m=%scan(&&months&i,&j.);
%end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);

%mend

%mnthly_bal_data

%macro mnthly_delinq_data;
%let k=24;
%let kminus4=%eval(&k-4);
%let kminus6=%eval(&k-6);

%put k &k. kminus4 &kminus4 kminus6 &kminus6;

%let i=1;
%let y=%scan(&year,&i);
%do %until ("&y"="");
    %let j=1;
    %let m=%scan(&&months&i,&j.);
%do %until ("&m"="");

data &bal_del..d;
set &input_lib..Mis_rolling(keep=ACCT_NUM DELQ_&k. STMTDTE_&k. DELQ_&kminus4. STMTDTE_&kminus4. DELQ_&kminus6. STMTDTE_&kminus6.);
rename DELQ_&k.=DELQ_now STMTDTE_&k.=DT_now DELQ_&kminus4.=DELQ_4_month STMTDTE_&kminus4.=DT_4_month DELQ_&kminus6.=DELQ_6_month STMTDTE_&kminus6.=DT_6_month;
run;

%let k=%eval(&k-1);
%let kminus4=%eval(&kminus4-1);
%let kminus6=%eval(&kminus6-1);

    %let j=%eval(&j+1);
    %let m=%scan(&&months&i,&j.);
%end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);

%mend

%mnthly_delinq_data;
Appendix B

SAS code to create Description and Category codes based on Merchant Category Codes provided by the bank

/******************************************************************************
 MERCHANT CATEGORY CODES
*******************************************************************************/

libname raw 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Raw_SAS_Data';
libname out 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Global_characteristics';
libname mccs 'C:\Documents and Settings\Administrator\Desktop\XXX_production\mcc_categories';
libname bal_del 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Balance_delinquency_data ';

/* get a list of all */

%let year= 06 07 08 ;
%let months1=10 11 12;
%let months2=01 02 03 04 05 06 07 08 09 10 11 12;
%let months3=01 02 03 04 05 06 07 08 09;

%macro generate_files;
    %let i=1;
    %let y=%scan(&year,&i);
    %do %until (&y="");
        %let j=1;
        %let m=%scan(&&months&i,&j.);
        %do %until ("&m"="");
            %let ym= &ym &y&m;
            %let filename = Txnlog_d&y&m;
            %put filename &filename.;
        %end;
    %end;
%end;

%macro generate_files;
    %let i=1;
    %let y=%scan(&year,&i);
    %do %until (&y="");
        %let j=1;
        %let m=%scan(&&months&i,&j.);
        %do %until ("&m"="");
            %let ym= &ym &y&m;
            %let filename = Txnlog_d&y&m;
            %put filename &filename.;
        %end;
    %end;
%end;

proc means data=raw.&filename. noprint;
var TRANSACTION_AMOUNT;
class MERCHANT CATEGORY_CODE;
output out=mccstats&y&m sum(TRANSACTION_AMOUNT)=total_spend
n(transaction_amount)=n;
run;

    %let j=%eval(&j+1);
%end;
%end;
%let y=%scan(&year,&i);
%end;
%mend;
%generate_files;

%macro append_files;
data mcc_stats(where=_type_=1);
set
  %let i=1;
  %let y=%scan(&year,&i);
%do %until ("&y"="");
    %let j=1;
    %let m=%scan(&&months&i,&j.);
    %do %until ("&m"="");
      %let ym= &ym &y&m;
      %put filename &filename.;
      mccstats&y&m
    %let j=%eval(&j+1);
    %let m=%scan(&&months&i,&j.);
  %end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);
%end;
run;
%mend;
%append_files;

proc means data=mcc_stats noprint;
var total_spend;
class MERCHANT_CATEGORY_CODE;
output out=mcc_stats_all sum(total_spend)=total_spend sum(n)=n;
run;

data mccs.mcc_stats_all;
set mcc_stats_all;
where _type_=1;
if MERCHANT_CATEGORY_CODE = . then delete;
drop _type_ _freq_;
run;

/* look at XXXs list of Merchant Codes*/

proc import
datafile="C:\Documents and Settings\Administrator\Desktop\XXX_production\mcc_categories\mcc.xls"
out=mccs.XXX_list_mcc
dbms=Excel2000
replace;
getnames=yes;
run;

proc sql;
create table mcc_lists as
select a.*, b.Description as XXX_desc
from mccs.mcc_stats_all as a left join mccs.XXX_list_mcc as b
on a.MERCHANT_CATEGORY_CODE = b.SDRETAILERSIC;
quit;

/* Check what % of mccs are allocated in list */
data missing_desc;
set mcc_lists;
where SDRETAILERSIC="";
run;

/* Get transactional descriptions to fill in missing XXX ones */
libname old 'C:\Documents and Settings\Administrator\Desktop\sas_heath';
data Mccs_desc_unique;
set mcc_lists;
mcc=MERCHANT_CATEGORY_CODE*1;
run;

/* Allocate merchant code descriptions without format */
data Farinaz_XXX_labels;
set Mccs_desc_unique;
attrib _all_ label="";
length desc $32 category $32;
if 3000<= mcc <=3299 or mcc in (4511,4582) then do
desc="AIRLINE";category="TRAVEL"; end;
if 3351<= mcc <=3441 or mcc in (7512,7513,7519) then do desc="CAR RENTAL";category="TRAVEL"; end;
if 3501<= mcc <=3814 or mcc in (2778,4411,6513,7011,7012,7033) then do
desc="LODGING";category="TRAVEL"; end;
if mcc in (4722,4723,5998,5598,5561,5309,5948,5962,5271) then do
desc="HOLIDAY_TRAVEL";category="TRAVEL"; end;
if mcc in (5977,5997,7230,7297,7298)
then do desc="HEALTH_BEAUTY";category="DISCRETIONARY"; end;
if 7832<= mcc <=7999 or mcc in (7032,7829,5996)
then do desc="SPORT_REC";category="DISCRETIONARY"; end;
if 5970<= mcc <=5972 or mcc in 
(4457,4468,5732,5733,5735,5945,5946,7221,7395)
then do desc="HOBBY";category="DISCRETIONARY"; end;
if 8398<= mcc <=8699
then do desc="DONATIONS";category="DISCRETIONARY"; end;
if 5611<= mcc <=5699 or mcc in (5699)
then do desc="CLOTHING";category="RETAIL"; end;
if 7210<= mcc <=7216 or mcc in (7251,7296,5137,5139,5949)
then do desc="CLOTHING";category="RETAIL"; end;
if 5310<= mcc <=5399 or mcc in 
(5192,5193,5300,5931,5933,5942,5944,5947,5963,5964,5965,
then do desc = "RETAIL"; category = "RETAIL"; end;
if 5712 <= mcc <= 5722 or mcc in (0780, 1799, 5932, 5937, 7934, 7623, 7629, 7217)
then do desc = "HOUSEHOLD_APPLIANCES"; category = "RETAIL"; end;
if 7622 <= mcc <= 7641 or mcc in (5950, 7699, 7349, 7394)
then do desc = "HOUSEHOLD_APPLIANCES"; category = "RETAIL"; end;
if 7372 <= mcc <= 7379 or mcc in (5734, 5021, 5044, 5045, 5072, 5111, 7333, 7399)
then do desc = "COMPUTER_COMMERCIAL"; category = "RETAIL"; end;
if 7523 <= mcc <= 7549 or mcc in (5013, 5172, 5571, 5599, 8675, 9752, 5983)
then do desc = "AUTOMOTIVE"; category = "TRANSPORT"; end;
if 5511 <= mcc <= 5542 then do desc = "AUTOMOTIVE"; category = "TRANSPORT"; end;
if mcc in (4011, 4012, 4111, 4112, 4121, 4131, 4789)
then do desc = "PUBLIC_TRANSPORT"; category = "TRANSPORT"; end;
if 5422 <= mcc <= 5499 or mcc in (5422)
then do desc = "FOOD"; category = "FOOD"; end;
if 5811 <= mcc <= 5814 or mcc in (5921, 5993)
then do desc = "EATING_DRINKING_OUT"; category = "FOOD"; end;
if mcc in (5411, 9751)
then do desc = "SUPERMARKET"; category = "FOOD"; end;
if 1711 <= mcc <= 1799 or mcc in (0763, 0780, 1799, 7692, 1520, 5074, 5065, 5198, 8911, 7342)
then do desc = "CONSTRUCTION_MAINTENANCE"; category = "CONSTRUCTION_MAINTENANCE"; end;
if 5200 <= mcc <= 5261 or mcc in (5039, 2842, 5051, 5085)
then do desc = "CONSTRUCTION_MAINTENANCE"; category = "CONSTRUCTION_MAINTENANCE"; end;
if mcc in (6010, 6011)
then do desc = "CASH_ADVANCE"; category = "CASH_ADVANCE"; end;
if 6012 <= mcc <= 6399 or mcc in (0000, 0744, 6529, 6530, 7276)
then do desc = "FINANCIAL_SERVICES"; category = "NON_DISCRETIONARY"; end;
if 4812 <= mcc <= 4821 or mcc in (4899, 4900, 9402, 9399)
then do desc = "UTILITIES"; category = "NON_DISCRETIONARY"; end;
if 8211 <= mcc <= 8299 or mcc in (5943, 8299)
then do desc = "EDUCATION"; category = "NON_DISCRETIONARY"; end;
if 8011 <= mcc <= 8099 or mcc in (4119, 5047, 5912, 5940, 5975, 5976, 5941, 8734, 5112, 0742)
then do desc = "MEDICAL"; category = "NON_DISCRETIONARY"; end;
if mcc in (0011, 0045, 2411, 5951)
then do desc = "SMALL_AMBIGUOUS";
category = "UNCLASSIFIED"; end;
if mcc in (5099, 5131, 5169, 5935, 7392)
then do desc = "NOT_RETAIL";
category = "UNCLASSIFIED"; end;
run;
data Mccs.Farinaz XXX_labels;
set Farinaz_XXX_labels;
if desc = "" then do desc = "UNCLASSIFIED";
category = "C_UNCLASSIFIED"; end;
if category = "UNCLASSIFIED" then do category = "C_UNCLASSIFIED"; end;
if category="NON_DISCRETIONARY" then do
category="C_NON_DISCRETIONARY";end;
if category="TRAVEL" then do category="C_TRAVEL";end;
if category="DISCRETIONARY" then do category="C_DISCRETIONARY";end;
if category="TRANSPORT" then do category="C_TRANSPORT";end;
if category="RETAIL" then do category="C_RETAIL";end;
if category="CONSTRUCTION_MAINTENANCE" then do
category="C_CONSTRUCTION_MAINTENANCE";end;
if category="FOOD" then do category="C_FOOD";end;
if category="CASH_ADVANCE" then do category="C_CASH_ADVANCE";end;

run;

data mccs.mcc_cat_desc;
set mccs.Farinaz_XXX_labels;
drop total_spend--mcc;
run;
Appendix C

SAS code to create AR parameters from daily time series

```sas
/********************************************************************
AR Parameters and Moments from Time series
********************************************************************/
/* Define Libraries */
libname raw 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Raw_SAS_Data';
libname out 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Global_characteristics';
libname mccs 'C:\Documents and Settings\Administrator\Desktop\XXX_production\mcc_categories';
libname bal_del 'C:\Documents and Settings\Administrator\Desktop\XXX_production\Balance_delinquency_data';

/* List years months of interest */
%let year= 07 08 ;
%let months1=03 04 05 06 07 08 09 10 11 12 ;
%let months2=01 02 03 04 05 06 07 08 09 ;

/* Point to libraries */

%let input_lib=raw;
%let output_lib=out;

/* Only keep positive transactions. Roll up transactional data to daily amounts */

/**/
/*%let y =07;*/
/*%let m =03;*/

%macro daily_trans;
%let i=1;
%let y=%scan(&year,&i);
%do %until ('&y'='''');
%let j=1;
%let m=%scan(&&months&i,&j.);
%do %until ('&m'='''');

%let filename = Txnlog_d&y&m; %put filename &filename.;

/* convert to sas date*/
data data;
  set &input_lib..&filename;
```

---

249
where transaction_amount > 0;
sasdate=mdy(substr(date_of_transaction, 5, 2), substr(date_of_transaction, 7, 2), substr(date_of_transaction, 1, 4));
keep merchant_category_code merchant_terminal_name transaction_amount transaction_code transaction_reason_code acct_num date_of_transaction sasdate;
run;

proc sort data=data;
by acct_num sasdate;
run;

data _null_; 
  month_end_sas=intnx('month', mdy(&m, 1, %eval(2000+&y)), 0, 'end');
  month_end=""||put(month_end_sas, date9.)||"d";
  call symput("month_end", month_end);

  month_start_sas=intnx('month', mdy(&m, 1, %eval(2000+&y)), 0, 'beginning');
  month_start=""||put(month_start_sas, date9.)||"d";
  call symput("month_start", month_start);
run;

proc timeseries data=data out=time_series_data&y&m 
MAXERROR=10;
  id sasdate interval=day accumulate=total SETMISSING=0
  START=&month_start 
  END=&month_end; 
  var transaction_amount;
  by acct_num;
run;

%let j=%eval(&j+1);
%let m=%scan(&&months&i,&j.);
%end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);
%end;
%mend;
%daily_trans;

/*for every series calculate moments and ARIMA parameters*/

%macro time_series_parameters;
%let i=1;
  %let y=%scan(&year,&i);
  %do %until ("&y"="")
    %let j=1;
    %let m=%scan(&&months&i,&j.);
    %do %until ("&m"="")

proc means data=time_series_data&y&m n mean skewness kurtosis nway noprint;
  var transaction_amount;
  class acct_num ;
output out=Input_moments&m(drop=_freq_ _FREQ_ _type_ _freq_
mean(transaction_amount)=mean
skewness(transaction_amount)=sk
kurtosis(transaction_amount)=kr
std(transaction_amount)=std;
run;
proc sql;
create table time_series_non_zero&m as
select * from time_series_data&m
where ACCT_NUM in (select ACCT_NUM from input_moments&m where std ne 0 and mean ne 0);
quit;
options nonotes;
proc arima data=time_series_non_zero&m;
by ACCT_NUM;
identify noprint var=transaction_amount alpha=.05 center p=(0:7)
q=(0:0);
estimate noprint p=7 outest=Input_arima_param&m(keep=_type_ ACCT_NUM MU--AR1_7 where=(_type_="EST"));
run;
options notes;

%let j=%eval(&j+1);
%let m=%scan(&&months&i,&j.);
%end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);
%end;
%mend;
%time_series_parameters;
Appendix D

SAS code for creation of input variables

/* Get % spend on each mcc description and category */
%macro monthly_mcc;
%let i=1;
%let y=%scan(&year,&i);
%do %until ("&y"="");
%let j=1;
%let m=%scan(&&months&i,&j.);
%do %until ("&m"="");
%let filename = Txnlog_d&y&m; %put filename
&filename.;
/* Get % spend on each mcc description and category */

/* get mcc categories for positive transactions */
proc sql;
create table data_mcc as
select a.merchant_category_code,
a.transaction_amount, a.acct_num,b.* /*, a.date_of_transaction*/
from data as a left join mccs.mcc_cat_desc as b
on a.MERCHANT_CATEGORY_CODE=b.MERCHANT_CATEGORY_CODE
where a.MERCHANT_CATEGORY_CODE ne ""
and b.MERCHANT_CATEGORY_CODE ne ""
and transaction_amount >0
order acct_num
;quit;
/* transaction spend on each description and category*/
proc freq data= data_mcc noprint;
weight transaction_amount;
by acct_num;
tables desc /out=mcc_desc_total;
tables category /out=mcc_category_total;
run;
/* % spend on each description */
proc transpose data=mcc_desc_total
out=mcc_trans_desc (drop= _name_);
var percent;
by acct_num;
*/
id desc;
run;

/* % spend on each category*/
proc transpose data=mcc_category_total
out=mcc_trans_cat(drop=_name_);
var percent;
by acct_num;
id category;
run;

/* join aggregated desc and category totals*/
proc sql;
create table Input_mcc_w_missing(drop=_label_)
as
select *
from  mcc_trans_cat as a inner join
mcc_trans_desc as b
on a.acct_num=b.acct_num
;quit;

/* Zero spend appears as a missing value after above
transpose, relace it here */
data Input_mcc&y&m;
set Input_mcc_w_missing;
array nvar(*) _numeric_;   
do i=1 to dim(nvar);
if nvar(i)=. then nvar(i)=0;
end;
drop i;
run;

%let j=%eval(&j+1);
%let m=%scan(&&months&i,&j.);
%end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);
%end;
%mend:  
%monthly_mcc;

/*generate payment variables */
%macro monthly_payments;
%let i=1;
%let y=%scan(&year,&i);
%do %until ("&y"="");
  %let j=1;
  %let m=%scan(&&months&i,&j.);
%do %until ("&m"="");
  %let filename = Txnlog_d&y&m;
  %put filename &filename.;
%end;
%end;
%mend:  
%monthly_mcc;
data data;
set &input_lib..&filename;
if transaction_amount < 0 then payment=transaction_amount;
else if transaction_amount > 0 then expenditure=transaction_amount;
keep acct_num expenditure payment;
run;

proc sort data=data;
by acct_num;
proc means data=data n mean max min noprint nway;
class acct_num;
output out=payments_expense(keep=acct_num monthly_expenditure monthly_payment_neg)
sum(expenditure)=monthly_expenditure sum(payment)=monthly_payment_neg;
run;

proc sql;
create table payments_expense_bal as
select *from payments_expense as a left join bal_del.Bal&y&m as b
on a.ACCT_NUM=b.ACCT_NUM;
quit;

data Input_payment_ratios&Y&M;
set payments_expense_bal;
if monthly_payment_neg=. then monthly_payment=0;
else monthly_payment=-monthly_payment_neg;
if monthly_expenditure=. then monthly_expenditure=0;
if monthly_expenditure+1 =0 then pay_spend_ratio =
monthly_payment/(monthly_expenditure);
else pay_spend_ratio = monthly_payment/(monthly_expenditure+1);
if STMTRBAL+1 = 0 then pay_bal_ratio = monthly_payment/(STMTRBAL);
else pay_bal_ratio = monthly_payment/(STMTRBAL+1);
if BALRLTL+BALCASH+1=0 then bal_change_bal_ratio= end_start_bal_ratio=
STMTRBAL/(BALRLTL+BALCASH);
else end_start_bal_ratio= STMTRBAL/(BALRLTL+BALCASH+1);
if BALRLTL+BALCASH+1 =0 then bal_change_bal_ratio= (STMTRBAL-
(BALRLTL+BALCASH))/(BALRLTL+BALCASH);
else bal_change_bal_ratio = (STMTRBAL-
(BALRLTL+BALCASH))/(BALRLTL+BALCASH+1);
drop STMTRDTE BALBT BALCASH STMTRBAL BALRLTL monthly_payment_neg
monthly_payment monthly_expenditure;
run;

%let j=%eval(&j+1);
%let m=%scan(&&months&i,&j.);
%end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);
%end;
%mend;
%monthly_payments;

/* merging all data to form input data*/

%macro logistic_input_data;
  %let i=1;
  %let y=%scan(&year,&i);
  %do %until (“&y”=””);
    %let j=1;
    %let m=%scan(&&months&i,&j.);
    %do %until (“&m”=””);
      /* Join all Global parameters: AR parameters and moments, mcc stats and payment ratios */
      proc sql;
      create table Input_Global_Params as select *from Input_moments&y&m as a inner join Input_arima_param&y&m(drop=_type_) as b on a.acct_num=b.acct_num inner join Input_mcc&y&m as c on a.acct_num=c.acct_num inner join Input_payment_ratios&Y&M as d on a.acct_num=d.acct_num;
      quit;

      /* Add balance and delinquency data */
      proc sql;
      create table &output_lib..Global_params_delinq_bal as select a.*, b.DELQ_now, b.DELQ_4_month, b.DELQ_6_month, b.dt_now, c.Bal_now, c.BAL_4_month, c.BAL_6_month, c.STMTDTE_NOW from Input_Global_Params as a inner join bal_del.Delinq&y&m as b on a.acct_num=b.acct_num inner join bal_del.Future_Bal&y&m as c on a.acct_num=c.acct_num; quit;

      /* Define modelling targets */
      data &output_lib..logistic_input&y&m;
      set &output_lib..Global_params_delinq_bal;
      where delq_now in (”0″,”1″);
      if DELQ_4_month in (″E″) then delete;
      if delq_4_month in (″L″) then delq_4_month=″6″;
      if DELQ_6_month in (″E″) then delete;
      if delq_6_month in (″L″) then delq_6_month=″6″;
      delq_4= input(strip(DELQ_4_month),8.);
      delq_6= input(strip(DELQ_6_month),8.);
      if DELQ_4 =0 then target_4m_delinq=0;
      if DELQ_4 ge 3 then target_4m_delinq=1;
      if DELQ_6 = 0 then target_6m_delinq=0;
      if DELQ_6 ge 3 then target_6m_delinq=1;
    %end;
  %end;
%mend;
rename BAL_4_month=target_4m_balance
BAL_6_month=target_6m_balance;

if target_4m_delinq=. then target_6m_delinq=.;

/* DELQ_4_month in ("1","2") are excluded by missing target */
/* "3" is three months delinquent, basel II definition */
/* sample 6m data from 4m data */

drop DELQ_now delq_4_month delq_6_month DT_now
STMTDTE_NOW delq_4 delq_6 BAL_now;
run;

/* Standardise all input variable for clustering */

PROC STANDARD DATA=&output_lib..logistic_input&y&m MEAN=0 STD=1
OUT=&output_lib..logistic_input_standardised&y&m;
VAR mean--end_start_bal_ratio;
RUN;

proc freq data=&output_lib..logistic_input&y&m;
table target_6m_delinq;
table target_4m_delinq;
run;

%let j=%eval(&j+1);
%let m=%scan(&&months&i,&j.);
%end;
%let i=%eval(&i+1);
%let y=%scan(&year,&i);
%end;
%mend;
%logistic_input_data;
## Appendix E

### E1. Table of Moments

<table>
<thead>
<tr>
<th>Name</th>
<th>Min</th>
<th>Max</th>
<th>Mean</th>
<th>StdDev</th>
<th>Skewness</th>
<th>Kurtosis</th>
</tr>
</thead>
<tbody>
<tr>
<td>MEAN</td>
<td>0.00</td>
<td>975.46</td>
<td>22.47</td>
<td>59.66</td>
<td>8.22</td>
<td>95.41</td>
</tr>
<tr>
<td>SK</td>
<td>1.14</td>
<td>7.81</td>
<td>4.70</td>
<td>1.13</td>
<td>-1.00</td>
<td>-0.28</td>
</tr>
<tr>
<td>KR</td>
<td>0.54</td>
<td>61.25</td>
<td>23.51</td>
<td>9.50</td>
<td>-0.80</td>
<td>-0.75</td>
</tr>
<tr>
<td>STD</td>
<td>0.02</td>
<td>5394.40</td>
<td>77.34</td>
<td>254.08</td>
<td>12.13</td>
<td>198.16</td>
</tr>
<tr>
<td>MU</td>
<td>-34.15</td>
<td>876.49</td>
<td>5.69</td>
<td>40.12</td>
<td>13.68</td>
<td>245.88</td>
</tr>
<tr>
<td>AR1_1</td>
<td>-0.50</td>
<td>1.32</td>
<td>0.01</td>
<td>0.18</td>
<td>3.33</td>
<td>13.95</td>
</tr>
<tr>
<td>AR1_2</td>
<td>-0.95</td>
<td>1.47</td>
<td>-0.03</td>
<td>0.16</td>
<td>1.20</td>
<td>13.98</td>
</tr>
<tr>
<td>AR1_3</td>
<td>-1.12</td>
<td>1.15</td>
<td>0.01</td>
<td>0.17</td>
<td>2.45</td>
<td>10.57</td>
</tr>
<tr>
<td>AR1_4</td>
<td>-0.89</td>
<td>0.96</td>
<td>-0.04</td>
<td>0.13</td>
<td>0.64</td>
<td>8.85</td>
</tr>
<tr>
<td>AR1_5</td>
<td>-0.50</td>
<td>0.78</td>
<td>-0.01</td>
<td>0.12</td>
<td>2.16</td>
<td>7.75</td>
</tr>
<tr>
<td>AR1_6</td>
<td>-0.69</td>
<td>0.82</td>
<td>-0.03</td>
<td>0.12</td>
<td>1.02</td>
<td>7.35</td>
</tr>
<tr>
<td>AR1_7</td>
<td>-0.97</td>
<td>0.91</td>
<td>-0.05</td>
<td>0.20</td>
<td>-2.55</td>
<td>12.60</td>
</tr>
<tr>
<td>C_NON_DISCRETIONARY</td>
<td>0.00</td>
<td>100.00</td>
<td>56.30</td>
<td>40.26</td>
<td>-0.09</td>
<td>-1.70</td>
</tr>
<tr>
<td>C_CONSTRUCTION_MAINTENANCE</td>
<td>0.00</td>
<td>97.86</td>
<td>2.45</td>
<td>11.04</td>
<td>6.16</td>
<td>41.58</td>
</tr>
<tr>
<td>C_FOOD</td>
<td>0.00</td>
<td>100.00</td>
<td>7.76</td>
<td>16.08</td>
<td>2.67</td>
<td>7.86</td>
</tr>
<tr>
<td>C_RETAIL</td>
<td>0.00</td>
<td>100.00</td>
<td>12.35</td>
<td>23.08</td>
<td>2.14</td>
<td>3.87</td>
</tr>
<tr>
<td>C_TRANSPORT</td>
<td>0.00</td>
<td>99.57</td>
<td>5.47</td>
<td>14.42</td>
<td>3.76</td>
<td>15.59</td>
</tr>
<tr>
<td>C_TRAVEL</td>
<td>0.00</td>
<td>99.97</td>
<td>4.06</td>
<td>14.47</td>
<td>4.24</td>
<td>18.54</td>
</tr>
<tr>
<td>C_UNCLASSIFIED</td>
<td>0.00</td>
<td>100.00</td>
<td>2.05</td>
<td>9.76</td>
<td>6.77</td>
<td>51.89</td>
</tr>
<tr>
<td>C_DISCRETIONARY</td>
<td>0.00</td>
<td>100.00</td>
<td>4.78</td>
<td>14.82</td>
<td>4.37</td>
<td>20.51</td>
</tr>
<tr>
<td>C_CASH_ADVANCE</td>
<td>0.00</td>
<td>99.87</td>
<td>4.78</td>
<td>17.23</td>
<td>3.84</td>
<td>14.02</td>
</tr>
<tr>
<td>FINANCIAL_SERVICES</td>
<td>0.00</td>
<td>100.00</td>
<td>44.91</td>
<td>43.71</td>
<td>0.32</td>
<td>-1.73</td>
</tr>
<tr>
<td>AUTOMOTIVE</td>
<td>0.00</td>
<td>99.57</td>
<td>5.13</td>
<td>13.98</td>
<td>3.93</td>
<td>17.11</td>
</tr>
<tr>
<td>CONSTRUCTION_MAINTENANCE</td>
<td>0.00</td>
<td>97.86</td>
<td>2.45</td>
<td>11.04</td>
<td>6.16</td>
<td>41.58</td>
</tr>
<tr>
<td>EATING_DRINKING_OUT</td>
<td>0.00</td>
<td>100.00</td>
<td>2.55</td>
<td>8.77</td>
<td>5.67</td>
<td>40.49</td>
</tr>
<tr>
<td>HOUSEHOLD_APPLIANCES</td>
<td>0.00</td>
<td>100.00</td>
<td>2.00</td>
<td>9.84</td>
<td>6.72</td>
<td>50.77</td>
</tr>
<tr>
<td>LODGING</td>
<td>0.00</td>
<td>99.26</td>
<td>1.69</td>
<td>8.75</td>
<td>6.93</td>
<td>53.84</td>
</tr>
<tr>
<td>SUPERMARKET</td>
<td>0.00</td>
<td>100.00</td>
<td>4.66</td>
<td>11.92</td>
<td>3.54</td>
<td>14.67</td>
</tr>
<tr>
<td>UTILITIES</td>
<td>0.00</td>
<td>100.00</td>
<td>6.36</td>
<td>17.05</td>
<td>3.43</td>
<td>12.26</td>
</tr>
<tr>
<td>FOOD</td>
<td>0.00</td>
<td>70.65</td>
<td>0.55</td>
<td>3.36</td>
<td>10.88</td>
<td>158.40</td>
</tr>
<tr>
<td>RETAIL</td>
<td>0.00</td>
<td>99.87</td>
<td>5.40</td>
<td>14.36</td>
<td>3.65</td>
<td>14.64</td>
</tr>
<tr>
<td>UNCLASSIFIED</td>
<td>0.00</td>
<td>100.00</td>
<td>1.95</td>
<td>9.59</td>
<td>6.98</td>
<td>55.20</td>
</tr>
<tr>
<td>AIRLINE</td>
<td>0.00</td>
<td>99.97</td>
<td>1.07</td>
<td>7.49</td>
<td>8.85</td>
<td>87.17</td>
</tr>
<tr>
<td>COMPUTER_COMMERCIAL</td>
<td>0.00</td>
<td>100.00</td>
<td>2.39</td>
<td>10.72</td>
<td>6.20</td>
<td>42.89</td>
</tr>
<tr>
<td>Name</td>
<td>Min</td>
<td>Max</td>
<td>Mean</td>
<td>StdDev</td>
<td>Skewness</td>
<td>Kurtosis</td>
</tr>
<tr>
<td>-----------------------</td>
<td>------</td>
<td>-------</td>
<td>-------</td>
<td>--------</td>
<td>----------</td>
<td>----------</td>
</tr>
<tr>
<td>DONATIONS</td>
<td>0.00</td>
<td>99.35</td>
<td>1.06</td>
<td>7.53</td>
<td>10.23</td>
<td>114.60</td>
</tr>
<tr>
<td>MEDICAL</td>
<td>0.00</td>
<td>100.00</td>
<td>4.19</td>
<td>13.71</td>
<td>4.55</td>
<td>22.68</td>
</tr>
<tr>
<td>CLOTHING</td>
<td>0.00</td>
<td>100.00</td>
<td>2.56</td>
<td>10.69</td>
<td>5.89</td>
<td>38.95</td>
</tr>
<tr>
<td>HEALTH_BEAUTY</td>
<td>0.00</td>
<td>94.79</td>
<td>0.94</td>
<td>6.24</td>
<td>10.76</td>
<td>133.68</td>
</tr>
<tr>
<td>NOT_RETAIL</td>
<td>0.00</td>
<td>56.90</td>
<td>0.09</td>
<td>1.87</td>
<td>27.34</td>
<td>793.42</td>
</tr>
<tr>
<td>PUBLIC_TRANSPORT</td>
<td>0.00</td>
<td>87.14</td>
<td>0.34</td>
<td>3.62</td>
<td>16.10</td>
<td>303.51</td>
</tr>
<tr>
<td>SPORT_REC</td>
<td>0.00</td>
<td>100.00</td>
<td>1.50</td>
<td>8.54</td>
<td>8.32</td>
<td>78.38</td>
</tr>
<tr>
<td>CASH_ADVANCE</td>
<td>0.00</td>
<td>99.87</td>
<td>4.78</td>
<td>17.23</td>
<td>3.84</td>
<td>14.02</td>
</tr>
<tr>
<td>EDUCATION</td>
<td>0.00</td>
<td>91.83</td>
<td>0.85</td>
<td>6.40</td>
<td>9.76</td>
<td>105.46</td>
</tr>
<tr>
<td>CAR_RENTAL</td>
<td>0.00</td>
<td>99.22</td>
<td>1.28</td>
<td>7.69</td>
<td>8.67</td>
<td>84.20</td>
</tr>
<tr>
<td>HOBBY</td>
<td>0.00</td>
<td>99.22</td>
<td>0.25</td>
<td>3.40</td>
<td>19.91</td>
<td>475.28</td>
</tr>
<tr>
<td>HOLIDAY_TRAVEL</td>
<td>0.00</td>
<td>95.80</td>
<td>1.05</td>
<td>7.69</td>
<td>8.88</td>
<td>85.39</td>
</tr>
<tr>
<td>PAY_SPEND_RATIO</td>
<td>0.00</td>
<td>626.96</td>
<td>3.71</td>
<td>25.04</td>
<td>15.75</td>
<td>314.36</td>
</tr>
<tr>
<td>PAY_BAL_RATIO</td>
<td>-191.80</td>
<td>32000.00</td>
<td>39.28</td>
<td>834.90</td>
<td>31.35</td>
<td>1117.00</td>
</tr>
<tr>
<td>BAL_CHANGE_BAL_RATIO</td>
<td>-742.00</td>
<td>8234.50</td>
<td>15.58</td>
<td>273.68</td>
<td>21.85</td>
<td>566.05</td>
</tr>
<tr>
<td>END_START_BAL_RATIO</td>
<td>-741.00</td>
<td>8235.20</td>
<td>16.51</td>
<td>273.63</td>
<td>21.86</td>
<td>566.32</td>
</tr>
<tr>
<td>BAL_NOW</td>
<td>-7346.0</td>
<td>34391.0</td>
<td>3914.5</td>
<td>5263.50</td>
<td>2.27</td>
<td>6.40</td>
</tr>
</tbody>
</table>
Appendix F

Table F1. Comparison of moments for few good and bad clusters

![Comparison of moments for few good and bad clusters](image)

Table F2. Comparison of percentage spent on different description codes for few good and bad clusters

![Comparison of percentage spent on different description codes](image)
Table F3. Comparison of Percentage spent on different categories for few good and bad clusters

<table>
<thead>
<tr>
<th>Cluster 33 (bad)</th>
<th>Cluster 21 (bad)</th>
<th>Cluster 4 (bad)</th>
<th>Cluster 3 (good)</th>
<th>Cluster 17 (good)</th>
<th>Cluster 30 (good)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
<td>Percentage</td>
</tr>
<tr>
<td>Construction, Maintenance</td>
<td>Food</td>
<td>Retail</td>
<td>Transport</td>
<td>Travel</td>
<td>Unclassified</td>
</tr>
<tr>
<td>0</td>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
</tr>
<tr>
<td>1</td>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
</tr>
<tr>
<td>2</td>
<td>3</td>
<td>4</td>
<td>5</td>
<td>6</td>
<td>7</td>
</tr>
</tbody>
</table>

---

261
Appendix G

Figure F1. ROC curve using Hybrid SOM based classification model using good/slow customers

![ROC Curve](image)

Gini Coefficient 0.69

Table F1. Number of good/slow and bad customers in March 2007 dataset.

<table>
<thead>
<tr>
<th></th>
<th>#</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>GOOD /SLOW</td>
<td>967477</td>
<td>99%</td>
</tr>
<tr>
<td>BAD</td>
<td>7171</td>
<td>1%</td>
</tr>
<tr>
<td>TOTAL</td>
<td>974648</td>
<td></td>
</tr>
</tbody>
</table>