2016-05-30

Investigation into the Predictive Capability of Macro-Economic Features in Modelling Credit Risk for Small Medium Enterprises

Kevin McTiernan
Technological University Dublin

Follow this and additional works at: https://arrow.tudublin.ie/scschcomdis

Part of the Computer Engineering Commons

Recommended Citation

This Theses, Masters is brought to you for free and open access by the School of Computing at ARROW@TU Dublin. It has been accepted for inclusion in Dissertations by an authorized administrator of ARROW@TU Dublin. For more information, please contact yvonne.desmond@tudublin.ie, arrow.admin@tudublin.ie, brian.widdis@tudublin.ie.

This work is licensed under a Creative Commons Attribution-Noncommercial-Share Alike 3.0 License
Dublin Institute of Technology

Masters Thesis

Investigation into the Predictive Capability of Macro-Economic Features in Modelling Credit Risk for Small Medium Enterprises

Author: Kevin McTiernan
Supervisor: Dr. Sarah Jane Delany

A dissertation submitted in partial fulfilment of the requirements of Dublin Institute of Technology for the degree of M.Sc. in Computing (Data Analytics)

March 2016
Declaration of Authorship

I, Kevin McTiernan, declare that this thesis titled, “Investigation into the Predictive Capability of Macro-Economic Features in Modelling Credit Risk for Small Medium Enterprises” and the work presented in it are my own. I confirm that:

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

Signed: 

Date:
“If we have data, let’s look at data. If all we have are opinions, let’s go with mine”

Jim Barksdale, former Netscape CEO
Abstract

This research project investigates the predictive capability of macro-economic features in modelling credit risk for small medium enterprises (SME/SMEs). There have been indications that there is strong correlation between economic growth and the size of the SME sector in an economy. However, since the financial crisis and consequent policies and regulations, SMEs have been hampered in attempts to access credit. It has also been noted that while there is a substantial amount of credit risk literature, there is little research on how macro-economic factors affect credit risk. Being able to improve credit scoring by even a small amount can have a very positive effect on a financial institution’s profits, reputation and ability to support the economy.

Typically, in the credit scoring process two methods of scoring are carried out, application scoring model and behavioural scoring model. These models for predicting customers who are likely to default usually rely upon financial, demographic and transactional data as the predictive inputs. This research investigates the use of a much coarser source of data at a macro-economic level by a low level and high level regions in Ireland. Features such as level of employment/unemployment, education attainment, consumer spending trends and default levels by different banking products will be evaluated as part of the research project.

In the course of this research, techniques and methods are established for evaluating the usefulness of macro-economic features. These are subsequently introduced into the predictive models to be evaluated. It was found that while employing coarse classification and subsequently choosing the macro-economic features with the highest information value in the predictive model, the accuracy across all performance measures improved significantly. This has proven that macro-economic features have the potential to be used in modelling credit risk for SMEs in the future.

Key words: Small medium enterprises, SME, credit risk, predictive modelling, macro-economic features
Acknowledgements

I would like to thank my supervisor, Dr. Sarah Jane Delany, for all the help she has provided to me in this project.
## Contents

Declaration of Authorship ............................................. i  
Acknowledgements ..................................................... iv  
Contents ........................................................................... v  
List of Figures ............................................................... viii  
List of Tables ................................................................. xi  
Abbreviations ................................................................... xiii  

1 Introduction
  1.1 Background ................................................................ 2  
  1.2 Research Project .................................................... 3  
  1.3 Research Objectives ................................................ 3  
  1.4 Research Methodology and Analytical Approach ............. 4  
  1.5 Scope and Limitations .............................................. 5  
  1.6 Outline of the Thesis ............................................... 6  

2 State-of-the-art
  2.1 Introduction ................................................................ 7  
  2.2 SME Definition ....................................................... 8  
  2.3 Credit Scoring ......................................................... 9  
  2.4 Importance of Credit Risk Modelling for SMEs ............... 11  
  2.5 Macro Area Features Affecting SME and Economic Credit Risk 12  
  2.6 Data Mining and Predictive Modelling ......................... 14  
  2.7 Dataset Construction .............................................. 17  
    2.7.1 Sampling Period .............................................. 18  
    2.7.2 Class Label Definition .................................... 18  
    2.7.3 Segmentation ................................................ 19  
  2.8 Predictive Models .................................................... 19  
    2.8.1 Regression ..................................................... 20  
    2.8.2 Logistic Regression .......................................... 21  
    2.8.3 K-Nearest Neighbour ...................................... 22  
    2.8.4 Decision Trees ............................................... 24  
    2.8.5 Artificial Neural Networks .................................. 27  
    2.8.6 Support Vector Machines ................................... 28
<table>
<thead>
<tr>
<th>Contents</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>A  Discretionary / Non-Discretionary Spend Features</td>
<td>122</td>
</tr>
<tr>
<td>B  Correlation Analysis</td>
<td>128</td>
</tr>
<tr>
<td>C  Correlation Analysis Filtered</td>
<td>131</td>
</tr>
<tr>
<td>D  Personal Loans and Homeloans Default Features</td>
<td>133</td>
</tr>
<tr>
<td>E  Central Statistics Office (CSO) Features</td>
<td>134</td>
</tr>
<tr>
<td>F  SME Default Behaviour</td>
<td>135</td>
</tr>
<tr>
<td>G  Feature Grouping Based on Home Loans, Personal Loans, SME Loans</td>
<td>136</td>
</tr>
</tbody>
</table>

Bibliography 137
List of Figures

2.1 Application Scorecard for Applicant X with a Credit Score of 355 ........................................ 11
2.3 Overview of the KDD process (Source: Fayyad et al., 1996) .................................................. 15
2.4 CRISP-DM Data Mining Process Model (Source: Shearer, 2000) ........................................... 15
2.5 Data Mining Venn Diagram ........................................................................................................ 16
2.6 Confusion Matrix ......................................................................................................................... 20
2.7 Standard Logistic Regression ...................................................................................................... 21
2.8 Comparison between Linear and Logistic Regression Models .................................................... 22
2.9 Example k-NN, Contrasting Results for $k=3$ and $k=5$ ............................................................... 23
2.10 Simple Decision Tree for Yes, No Prediction (Source: Quinlan, 1986) .................................. 24
2.11 Decision Tree with Nodes and Leaves labelled .......................................................................... 24
2.12 Simple and Complex Decision Tree Comparison (Source: Quinlan, 1986) .......................... 26
2.13 Three-layered Feed-forward Artificial Neural Network Configuration (Source: Raju et al., 2011) ........................................................................................................... 28
2.14 Optimal Separating Hyperplane in SVMs of Feature Space (Source: Li et al., 2011) ................. 29
2.15 SVM Classifying Non-Linearly Classes (Source: Burges, 1998) ............................................. 29
2.16 Impact of Absolute Rarity in Data Mining (Source: Weiss, 2004) ........................................ 34
2.17 Example of the K-NN for $x_i$ using $k = 6$. Data created using SMOTE based on the Euclidean distance. (Source: He and Garcia, 2009) ............................. 36
2.18 Example of Holdout Data Split with Training and Test Data ..................................................... 38
2.19 5-Fold Cross Validation ............................................................................................................. 38
2.20 Confusion Matrix ....................................................................................................................... 41
2.21 Confusion Matrix Example .......................................................................................................... 42
2.22 ROC ........................................................................................................................................ 43
2.23 Confusion Matrix: Multiple Threshold Comparison .................................................................... 44
3.1 Experiment Performance Window and Outcome Window .......................................................... 48
3.2 Republic of Ireland’s Electoral Divisions Local Authorities ....................................................... 51
3.3 Optimal Entity Relationship Data Model for Mapping Customers, Addresses, Electoral Division and Local Authorities .......................................................... 52
3.4 Current Entity Relationship Data Model for Mapping Customers, Addresses, Electoral Division and Local Authorities ......................................................... 53
3.5 GeoDirectory Database Diagram ................................................................................................ 54
3.6 Illustration of Inverted Indexing .................................................................................................. 55
3.7 Solr Query Example and Syntax ................................................................................................... 56
3.8 Address Matching Application for this Experiment .................................................................... 57
<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.9</td>
<td>Default Ratios for Homeloan and Personal Loans by Local Authority</td>
<td>62</td>
</tr>
<tr>
<td>3.10</td>
<td>ABT Generation Process</td>
<td>67</td>
</tr>
<tr>
<td>4.1</td>
<td>Overview of Benchmark Evaluation Process for Model and Threshold Selection</td>
<td>74</td>
</tr>
<tr>
<td>4.2</td>
<td>Previous Delinquency Benchmark Model ROC Charts</td>
<td>75</td>
</tr>
<tr>
<td>4.3</td>
<td>No Previous Delinquency Model ROC Chart</td>
<td>79</td>
</tr>
<tr>
<td>4.4</td>
<td>Overview of Experiment to Evaluate Macro-Economic Features for Previous Delinquency Dataset and No Previous Delinquency Dataset</td>
<td>81</td>
</tr>
<tr>
<td>5.1</td>
<td>SME Customers Default/Performing June 2012 - June 2015 x-axis = Year-Month y-axis = # SME Customers Red Trend Line = # SME Customers in Default each month Blue Trend Line = # SME Customers in Performing each month</td>
<td>86</td>
</tr>
<tr>
<td>5.2</td>
<td>Percentage of SME Customers in Default Each Month June 2012 - June 2015 x-axis = Year-Month y-axis = % of SME Customers in Default Black Trend Line = % SME Customers in Default Each Month</td>
<td>86</td>
</tr>
<tr>
<td>5.3</td>
<td>% SME Customers in Default by 34 Local Authority and 3,440 Electoral Division</td>
<td>87</td>
</tr>
<tr>
<td>5.4</td>
<td>Correlation Analysis of Macro-Economic Features</td>
<td>91</td>
</tr>
<tr>
<td>5.5</td>
<td>Correlation Analysis after Highly Correlated Features removal</td>
<td>92</td>
</tr>
<tr>
<td>5.6</td>
<td>Most correlated features with target feature</td>
<td>93</td>
</tr>
<tr>
<td>5.7</td>
<td>Top 20 Macro-Economic Feature Calculated by Information Gain on Previous Delinquency Dataset</td>
<td>96</td>
</tr>
<tr>
<td>5.8</td>
<td>Top 20 Macro-Economic Feature Calculated by Information Gain on No Previous Delinquency Dataset</td>
<td>98</td>
</tr>
<tr>
<td>5.9</td>
<td>Top 20 Macro-Economic Feature Calculated by Random Forest Feature Importance on Previous Delinquency Dataset</td>
<td>101</td>
</tr>
<tr>
<td>5.10</td>
<td>Top 20 Macro-Economic Feature Calculated by Random Forest Feature Importance on No Previous Delinquency Dataset</td>
<td>103</td>
</tr>
<tr>
<td>5.11</td>
<td>Feature Ranking by Information Value after Coarse Classification is Applied to the Macro-Economic Features of the Previous Delinquency Dataset</td>
<td>105</td>
</tr>
<tr>
<td>5.12</td>
<td>Barplot of Feature Ranking by Information Value after Coarse Classification is Applied to the Macro-Economic Features of the Previous Delinquency Dataset</td>
<td>106</td>
</tr>
<tr>
<td>5.13</td>
<td>Top 5 Binned Features Results based on Information Value after Coarse Classification is Applied to the Macro-economic Features of the Previous Delinquency Dataset</td>
<td>107</td>
</tr>
<tr>
<td>5.14</td>
<td>Feature ranking by Information Value after Coarse Classification is Applied to the Macro-economic Features of the No Previous Delinquency Dataset</td>
<td>109</td>
</tr>
<tr>
<td>5.15</td>
<td>Barplot of Feature Ranking by Information Value after Coarse Classification is Applied to the Macro-economic Features of the No Previous Delinquency Dataset</td>
<td>110</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>B.1</td>
<td>Grouped Features Correlation Analysis</td>
<td>128</td>
</tr>
<tr>
<td>B.2</td>
<td>SME Arrears Trends Correlation Analysis</td>
<td>129</td>
</tr>
<tr>
<td>B.3</td>
<td>Transaction Visa Debit Correlation Analysis</td>
<td>130</td>
</tr>
</tbody>
</table>
C.1 Grouped Features Correlation Analysis after Highly Correlated Features have been Removed ................................................. 131
C.2 SME Arrears Trends Correlation Analysis after Highly Correlated Features have been Removed ................................................. 132
C.3 Transaction Visa Debit Correlation Analysis after Highly Correlated Features have been Removed ................................................. 132
List of Tables

3.1 Characteristics of datasets to be used in the exploratory evaluation for training a benchmark model and assessing the evaluation metrics to be used in the research # Numeric refers to the number of continuous features # Nominal refers to the number of categorical features ............... 50
3.2 MCC Categories and Spend/Live Categorisation Sample .......................... 59
3.3 AIB Personal Loan and Homeloan products ......................................... 61
3.4 Personal Customers Arrears Ratio features derived for experiment .......... 62
3.5 Themes of data available in the Irish census 2011 ............................... 63
3.6 (CSO) Features derived for experiment ............................................. 64
3.7 SME default behaviour features derived for experiment ....................... 65
3.8 Binned/Grouped features based on default ratios of SME, Home and Personal Loans by Electoral Division and Local Authority ....................... 66
3.9 Characteristics of datasets to be used in the experiment to evaluate macro-economic features # Numeric refers to the number of continuous features # Nominal refers to the number of categorical features ............... 67

4.1 Breakdown Holdout Training/Validation/Test Dataset for Benchmark Evaluation ................................................................. 73
4.2 Previous Delinquency Benchmark Results ......................................... 75
4.3 Previous Delinquency Cut-off Results K-S = Optimal threshold for Kologorov-Smirnov Statistic Min MR = Optimal threshold for Min Mis-classification Rate EPER = Optimal threshold for Event Precision Equals Recall BA = Balanced Accuracy = \( \frac{\text{Recall} + \text{Specificity}}{2} \) ........................................... 77
4.4 No Previous Delinquency Bench Mark Results ..................................... 78
4.5 No Previous Delinquency Cut-off Results K-S = Optimal threshold for Kologorov-Smirnov Statistic MR = Optimal threshold for Mis-classification Rate EPER = Optimal threshold for Event Precision Equals Recall BA = Balanced Accuracy = \( \frac{\text{Recall} + \text{Specificity}}{2} \) ........................................... 80

5.1 Breakdown Holdout Training/Test Dataset for Benchmark Models ........ 89
5.2 Benchmark Model Results for Experiment Comparison ........................ 89
5.3 Previous Delinquency Model results when most important Macro-economic features calculated using correlation were included in training. *BA = Balanced Accuracy .................................................... 93
5.4 No Previous Delinquency Model results when most important Macro-economic features calculated using correlation were included in training. *BA = Balanced Accuracy .................................................... 94
<table>
<thead>
<tr>
<th>Table Number</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>5.5</td>
<td>Previous Delinquency Model results when most important Macro-economic features calculated using Information Gain were included in training. *BA = Balanced Accuracy</td>
<td>97</td>
</tr>
<tr>
<td>5.6</td>
<td>No Previous Delinquency Model results when most important Macro-economic features calculated using Information Gain were included in training. *BA = Balanced Accuracy</td>
<td>99</td>
</tr>
<tr>
<td>5.7</td>
<td>Previous Delinquency Model results when most important Macro-economic features calculated using Random Forest feature selection were included in training. *BA = Balanced Accuracy</td>
<td>102</td>
</tr>
<tr>
<td>5.8</td>
<td>No Previous Delinquency Model results when most important Macro-economic features calculated using Random Forest feature selection were included in training. *BA = Balanced Accuracy</td>
<td>104</td>
</tr>
<tr>
<td>5.9</td>
<td>Previous Delinquency Model results when most important Macro-economic features calculated using coarse classification were included. *BA = Balanced Accuracy</td>
<td>108</td>
</tr>
<tr>
<td>5.10</td>
<td>No Previous Delinquency Model results when most important Macro-Economic features calculated using coarse classification were included. *BA = Balanced Accuracy</td>
<td>110</td>
</tr>
<tr>
<td>5.11</td>
<td>Breakdown Holdout Training/Test Dataset for oversampled previous delinquency models</td>
<td>113</td>
</tr>
<tr>
<td>5.12</td>
<td>Oversampled Previous Delinquency Model results when most important Macro-economic features from PD_RF10_R model are included. *BA = Balanced Accuracy</td>
<td>113</td>
</tr>
<tr>
<td>5.13</td>
<td>Breakdown Holdout Training/Test Dataset for Oversampled No Previous Delinquency Models</td>
<td>114</td>
</tr>
<tr>
<td>5.14</td>
<td>Oversampled No Previous Delinquency Model results when most important Macro-economic features from NPD_IG5_R model were included. *BA = Balanced Accuracy</td>
<td>114</td>
</tr>
<tr>
<td>A.1</td>
<td>Discretionary Spend and Non-Discretionary Spend features derived for experiment</td>
<td>127</td>
</tr>
<tr>
<td>D.1</td>
<td>Personal Customers Arrears Ratio features derived for experiment</td>
<td>133</td>
</tr>
<tr>
<td>E.1</td>
<td>(CSO) Features derived for experiment</td>
<td>134</td>
</tr>
<tr>
<td>F.1</td>
<td>SME default behaviour features derived for experiment</td>
<td>135</td>
</tr>
<tr>
<td>G.1</td>
<td>Binned/Grouped features based on default ratios of SME, Home and Personal Loans by Electoral Division and Local Authority</td>
<td>136</td>
</tr>
</tbody>
</table>
## Abbreviations

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Full Form</th>
</tr>
</thead>
<tbody>
<tr>
<td>ABT</td>
<td>Analytics Base Table</td>
</tr>
<tr>
<td>AIB</td>
<td>Allied Irish Banks</td>
</tr>
<tr>
<td>ANN</td>
<td>Artificial Neural Network</td>
</tr>
<tr>
<td>API</td>
<td>Application Programming Interface</td>
</tr>
<tr>
<td>AUC</td>
<td>Area Under Curve</td>
</tr>
<tr>
<td>BA</td>
<td>Balanced Accuracy</td>
</tr>
<tr>
<td>CRAN</td>
<td>The Comprehensive R Archive Network</td>
</tr>
<tr>
<td>CRISP-DM</td>
<td>Cross Industry Standard Process for Data Mining</td>
</tr>
<tr>
<td>CSO</td>
<td>Central Statistics Office</td>
</tr>
<tr>
<td>ED</td>
<td>Electoral Division</td>
</tr>
<tr>
<td>EDW</td>
<td>Enterprise Data Warehouse</td>
</tr>
<tr>
<td>EPER</td>
<td>Event Precision Equals Recall</td>
</tr>
<tr>
<td>FN</td>
<td>False Negative</td>
</tr>
<tr>
<td>FP</td>
<td>False Positive</td>
</tr>
<tr>
<td>GDD</td>
<td>Geo Directory Database</td>
</tr>
<tr>
<td>GIS</td>
<td>Geographic Information System</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning System</td>
</tr>
<tr>
<td>ID3</td>
<td>Iterative Dichotomiser 3</td>
</tr>
<tr>
<td>IG</td>
<td>Information Gain</td>
</tr>
<tr>
<td>IV</td>
<td>Information Value</td>
</tr>
<tr>
<td>KDD</td>
<td>Knowledge Discovery in Databases</td>
</tr>
<tr>
<td>KNN</td>
<td>K Nearest in Neighbours</td>
</tr>
<tr>
<td>K-S</td>
<td>Kolmogorov Smirnov</td>
</tr>
<tr>
<td>LA</td>
<td>Local Authority</td>
</tr>
<tr>
<td>LOOCV</td>
<td>Leave One Out Cross Validation</td>
</tr>
<tr>
<td>Abbreviation</td>
<td>Description</td>
</tr>
<tr>
<td>--------------</td>
<td>-------------</td>
</tr>
<tr>
<td>MCC</td>
<td>Merchant Category Code</td>
</tr>
<tr>
<td>MR</td>
<td>Misclassification Rate</td>
</tr>
<tr>
<td>OECD</td>
<td>The Organisation for Economic Co-operation and Development</td>
</tr>
<tr>
<td>R</td>
<td>The R Project for Statistical Computing</td>
</tr>
<tr>
<td>ROC</td>
<td>Receiver Operating Characteristic</td>
</tr>
<tr>
<td>TN</td>
<td>True Negative</td>
</tr>
<tr>
<td>TP</td>
<td>True Positive</td>
</tr>
<tr>
<td>SAS</td>
<td>Statistical Analysis System</td>
</tr>
<tr>
<td>SME/SMEs</td>
<td>Small and Medium-sized Enterprises</td>
</tr>
<tr>
<td>SMOTE</td>
<td>Synthetic Minority Oversampling Technique</td>
</tr>
<tr>
<td>SQL</td>
<td>Structured Query Language</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machines</td>
</tr>
<tr>
<td>WoE</td>
<td>Weight of Evidence</td>
</tr>
</tbody>
</table>
Chapter 1

Introduction

Overview of Project Area

In finance, credit risk is one of the oldest forms of risk. Credit risk can be defined as one party, “a lender” trusting another party “a borrower”, enough that they are happy to give money “some credit” which they anticipate will be paid back not instantly but after some time interval. The credit risk is the probability or chance that the “lender” is never paid back by the “borrower”. Since lending commenced in 1800 B.C., there has existed a certain amount of uncertainty in the lending-borrowing process (Caouette et al., 1998).

There are many credit risks in the repayment process whereby the lender may not receive the full payment, the principal, the interest or anything at all. To help mitigate this issue, financial institutions have employed credit risk scoring and modelling to identify those borrowers likely to fail in paying back on their financial obligation (Srirattanaphonkun and Pattarathammamas, 2012).

The phrase credit scoring is commonly used to define the procedure of assessing the risk a borrower poses of defaulting on their financial agreement (Hand and Henley, 1997). The aim of these models is to classify borrowers “customers of the financial institution” into one of two classes: good and bad. Good and bad can also be referred to as defaulters or non-defaulters/performing. Customers in the good class are thought to be more likely to pay back their financial agreement while customers in the bad classes are thought to be unlikely to pay back their financial agreement.
Small and Medium-Sized Enterprises (SME/SMEs) are commonly defined as registered businesses with fewer than 250 employees (IFC, 2009). It should be noted however, that there is no consistent definition due to the variance across countries and financial institutions. In various research studies, it emerges that commercial financial institutions consider the SME sector to be very profitable (Beck et al., 2008). Additionally research demonstrates that performing SMEs or a strong SME sector is important as it forms the spine of countries’ economies around the world. This is because countries with a strong SME sector are hubs for providing jobs, innovation and growth to the economy (Craig et al., 2004). It is therefore vital from the perspective of both a financial institution’s profitability and a country’s economy, that credit risk models are as accurate as possible and have all the relevant information available to make informed predictions.

Since the financial crisis of 2007-2008 there has been a much greater emphasis on credit scoring for the entire consumer lending process in financial institutions. One of the most common methods of building credit risk models is by using data mining (Baesens et al., 2009). A key benefit to improving the scoring accuracy of a credit model is the significant future savings (West, 2000) but financial institutions are also under increasing regulatory pressures from global (Bank for International Settlements) e.g. The European Central Bank and National bank bodies e.g. Central Bank of Ireland. Since the crisis, these regulators actively police the dealings of financial institutions to ensure better care of their credit scoring systems. Poor performing credit systems can have massive adverse effects on financial institutions’ profits, reputations and ability to support the economy.

1.1 Background

Credit scoring processes carried out by financial institutions can generally be split into two groups (Bijak and Thomas, 2012). These processes will differ in terms of what data is used for scoring and the task they are trying to perform.

Firstly, application scoring, is employed when an application for credit is submitted. The application credit scoring model evaluates an applicant’s probability of defaulting at a later point in time based on the applicant’s credit application details. Financial and demographic information are typically used for this model where the current application
details are compared against previous applications with the same features along with their good/bad state at a later point in time.

Secondly, *behavioural scoring*, is employed once the borrower has secured credit from the lender. The *behavioural credit scoring model* evaluates the borrower’s probability of defaulting at a later point in time once the borrower has secured credit. This allows financial institutions to monitor constantly the performance of the borrower performance, enabling them to aid them if they are seen to be showing signs of *financial stress*. The predictive features that are typically used building this model are commonly based on borrower’s lending repayment performance and the borrower’s good/bad classification at some time in the future.

If financial institutions want to be sustainable and profitable it is imperative that they are able to identify accurately those borrowers who are likely to default in the future. For borrowers found to be of high risk, it allows the financial institution to make suitable decisions to mitigate the impact from its losses. The experiments in this research paper will focus only on the *behavioural scoring* aspect of *credit scoring*.

### 1.2 Research Project

The aim of this research project is to generate macro-economic features and assess their capability in predicting SME customers that will default on their financial obligation in AIB in the future.

### 1.3 Research Objectives

The primary goal of this research is to assess the predictive capability of macro-economic features in predicting whether or not SMEs will default on their financial obligation. The predictive models built as part of the experiment will include macro-economic features that will be sourced from internal sources in AIB and open datasets from the Irish Census.

The objectives of this research are:
To study the relevant state-of-the-art literature and industry best practices for credit risk scoring, predictive modelling and how macro-economic features are utilised in credit risk modelling.

Design and build an application to generate, collate and identify macro-economic features that will be assessed for predicting SMEs potential for default.

Design experiments to test the hypothesis.

Use feature selection techniques to identify the most predictive macro-economic features.

Train benchmark predictive models to compare and evaluate the experiment models.

Train predictive models including macro-economic features to be evaluated against the benchmark models.

Critically assess the results from predictive models including macro-economic features compared to the benchmark model to evaluate if macro-economic features should be included in credit risk models in the future.

Determine what future research could be undertaken in the area to expand on the research project.

1.4 Research Methodology and Analytical Approach

The research methodology that will be deployed in this project is empirical evaluation, involving investigation of and experimentation on a large number of macro-economic features. These features will be generated based on customer transactional spending behaviour, default trends over many banking credit products (personal loans, homeloans, SMEs) and Census data (employment levels, education levels, and occupation types).

The experiment undertaken in this research is based on building a prediction model that is able to predict accurately if SME customers will default or not in the future. As part of the experiment, macro-economic features are evaluated to ascertain if any are accurately able to predict arrears. Building prediction models and evaluating the prediction power of features are common practices in Data mining.
Data mining is used to explain historic events and forecast future events by applying data analysis. Data mining is commonly used to identify trends that are not immediately obvious.

In this experiment, the aim is to investigate the relationship between good and bad SME customers by including macro-economic features by region to see if there are any relationships from the past that could be used to predict the future. It is not feasible to investigate all these relationships manually. Data mining gives one a methodology which is supported by data analysis and experiments.

Data mining techniques for measuring the importance of features and evaluating the performance of prediction models will be used throughout the empirical evaluation of this project. Many tests will be carried out to try and derive interesting insights from the macro-economic features relationship with SME customer default behaviour.

1.5 Scope and Limitations

The scope of this project is to build a prediction model for SMEs in AIB which utilises macro-economic features by geographic regions (Electoral Division and Local Authority) in the Republic of Ireland. The aim of the experiment is to evaluate the predictiveness of these macro-economic features and evaluate if they should be included in industry credit risk models in AIB in the future.

SMEs included in this project will only be taken from one of the accounting systems in AIB. Macro-economic features for this experiment will be generated from customer transactional spending behaviour, default trends over disparate banking credit products (personal loans, homeloans, SMEs) and Census data (employment levels, education levels, and occupation types).

As part of the experiment, a benchmark prediction model will be trained using features that were selected to be in a SME credit risk model in the past. Further prediction models will be built using the features from the benchmark model and the macro-economic features generated as part of this research. Prediction models using macro-economic features will be compared and evaluated against the benchmark model. If the models trained using the macro-economic features perform better than the benchmark
model, then these features should be considered for inclusion in the SME credit risk model in the future.

1.6 Outline of the Thesis

The remaining chapters of this thesis are organised as follows:

- Chapter 2 documents and evaluates the current state of the art in the field of credit scoring, the use of macro-economic features in credit scoring, and the general field of data mining which includes predictive modelling, performance measurement and handling imbalanced datasets. Techniques and methods for feature set reduction and selection are also discussed here.

- Chapter 3 presents the data that will be used for the experiments in this research. It will detail what SME customers are used for the experiment, where they were sourced, under what criteria and what period of time the experiment will be conducted. It will also include details of how the macro-economic features for this experiment were generated. This will include details of where the data is sourced and details of how this data was mapped to geographic regions.

- Chapter 4 presents the design and research methodology for the project in an attempt to improve upon the prediction of a benchmark predictive model by introducing macro-economic features. Feature selection and performance measurement techniques for the experiment will addressed in this chapter.

- Chapter 5 presents the implementation of the experiments carried out as part of this research. Results will be evaluated and critically assessed. Conclusions and observations will be made where it is possible to do so.

- Chapter 6 concludes this thesis paper by summarising the contributions made to the problem of modelling credit risk for SMEs using macro-economic features. It concludes by discussing future research that could be carried out in this field and some alternative experiments worth implementing.
Chapter 2

State-of-the-art

2.1 Introduction

This chapter will discuss the research literature in the field of credit risk and default prediction of Small Medium Enterprises (SME/SMEs) customers in financial institutions. The first sections will cover off the areas of SME, detailing SME definitions, credit risk and how macro-economic features are utilised in modelling credit risk for SMEs and economies. These early sections will discuss challenges observed in the field such as a lack of shared definitions and statistics while it also details recommendations made to strengthen the field. The literature for macro-economic features is reviewed and it is noted that there is a dearth of research how macro-economic factors affect SME credit risk specifically and credit risk as a whole. The review concludes by detailing successful examples Italy and Portugal of applying macro-economic factors in predicting credit risk by utilising features based on default rates and unemployment by geographic location.

The chapter will also review research literature in the field of knowledge discovery, data mining with a particular focus on predictive modelling. Knowledge discovery and data mining will be explained and illustrated with frameworks and methodologies around the approach to tasks in each. A review of the literature facilitates an understanding of the processes, prediction algorithms, feature selection methods, validation methods and performance measures required to build a predictive model to predict SME customers that are likely to default i.e. risky customers. A key observation emerging from building a model to predict what customers will default is the tenancy towards a very large
class imbalance in the dataset e.g. there will be a much larger proportion of good/well performing than bad/poorly performing customers. Methods of addressing this class imbalance are also discussed later in this chapter.

### 2.2 SME Definition

The most common definition for a SME is a registered business with fewer than 250 employees (IFC, 2009). However this definition is not universally agreed: there are variances in the definition between countries and even across financial institutions.

At a European level a SME business is categorised as SME if they have two hundred and fifty people or fewer employed and if the annual turnover does not surpass €50 million, and/or an annual balance sheet not surpassing €43 million\(^1\). SMEs can also be subdivided further into smaller subcategories. Micro enterprises are defined as businesses that employ fewer than ten people, have annual turnover below €2 million, and annual balance sheet total not surpassing €2 million. Small enterprises are defined as businesses employing between ten and fifty people, have annual turnover below €10 million, and annual balance sheet total below €10 million. Medium enterprises are defined as businesses with an employee number of between fifty and two hundred and fifty people, have an annual turnover less than €50 million, and an annual balance sheet total below €43 million\(^2\).

Worldwide, the most common method by regulators for defining businesses as SME are based on the number of people they have employed, sales/turnover or/and loan size (Ardic et al., 2011).

In 2004, at the Organisation for Economic Co-operation and Development (OECD) conference on SMEs, two key recommendations were made by member economies and non-member states\(^3\): (i) develop SME statistics that can be compared internationally, and (ii) establish a common definition and set of rules for what is a SME. Without these statistics and definitions in place it would be more difficult to deploy programmes aiming to expand and strengthen the SME sector (Ardic et al., 2011). The aim of these

---

3. [http://www.cgap.org/financialindicators](http://www.cgap.org/financialindicators)
recommendations is that, by having a consistent definition and statistics for SMEs, economies and analysts can learn from each other on a global scale with an end goal of building a stronger SME sector.

2.3 Credit Scoring

Financial institutions use classification systems called credit scoring to evaluate the credit risks related with lending to a borrower. Credit risk is the risk of losses expected when a borrower’s ability or willingness to repay a financial obligation is adversely affected (Anderson, 2007). Credit scoring is the phrase used to encompass the methods and prediction techniques used by lenders to assess the credit risk of existing and prospective borrowers.

The aim of credit scoring is to classify prospective borrowers and existing borrowers into one of two groups, good or bad. The bad group signifies a borrower who was deemed likely to default on the financial obligation. The good group signifies borrowers deemed likely to repay their financial obligation.

Credit scoring models are typically broken into two main categories, application scoring and behavioural scoring. The objective of the application scoring model is to predict at the time the application is made, the borrower’s probability of defaulting at some time in the future. Application, product and demographic details are generally used to build the application scoring model. The objective of the behavioural scoring model is to predict the probability of existing customers defaulting. The borrower’s repayment performance is mainly used to build the behavioural scoring model.

Before credit scoring systems were employed by financial institutions, the risk of a borrower was based on the biased opinion of lender who would call on their life and work experience. Information about borrowers was gathered through personal relationships between borrowers and the employees of the lender (Anderson, 2007). Further investigation was carried out by means of a process known as the 5Cs

(i) **Character** - does the borrower, or their family, have a relationship with the lender?

(ii) **Capital** - what is credit amount requested?
(iii) **Collateral** - is security being offered?

(iv) **Capacity** - how fit is the borrower to repay?

(v) **Condition** - how is the economy performing presently?

This process was clearly flawed as it would not offer the financial institutions any consistency or reliability in terms of to which borrowers it was lending. With large improvements in computer power in the 1980s financial institutions started utilising analytical methods to gain a deeper understanding of customer behaviour (Hand, 2001).

Credit scoring systems are imperfect solutions as they can only be used as estimates based on historic events or the past's events, but not future events. A significant amount of debt goes unpaid each year due to the failure of credit scoring systems to identify borrowers who will default on their financial obligation (Finlay, 2011).

Generally a credit scoring system is built using a credit scorecard. Scorecard points are added based on important borrower characteristics and a score is generated that represents the risk of that borrower relative to all the other borrowers, in order of who is most likely to default on their financial obligation.

The most basic credit scorecards consist of a set of features that are statistically supported to be predicting the credit risk of a borrower (Siddiqi, 2012). One of the reasons for building a credit scorecard is for financial institutions to have a standardised, structured and easy to interpret mechanism of assessing borrower’s credit worthiness.

The credit scorecard in Fig. 2.1 uses features such as age, previous banking history, credit card limits, years at current job, accommodation status, self-employed status and monthly income to assign Applicant X a credit score. It can be seen in Fig. 2.1 that each feature is split into attributes and for each attribute a score has been generated which is added to the overall credit score for that applicant.
It can be observed from Fig. 2.1 that high value attributes are associated with borrowers that are statistically thought to be less likely to default on a financial obligation. For example, Applicant X in the example had an overall score of 355; if there were another applicant, Applicant Y, who had an overall score of 450, Applicant Y would score higher and therefore less likely to default on the financial obligation than Applicant X.

### 2.4 Importance of Credit Risk Modelling for SMEs

Recent research indicates that SME development is closely linked with economic growth. Beck et al. (2005) indicates there are strong correlations between economic growth and the size of the SME sector in an economy. After the recent global financial crisis of 2008 and 2009 it is important to note that the SME segment is recognised as one of the most
important factors contributing to sustainable employment and economic recovery (Lawless, McCann, McIndoe-Calder and others, 2012). Lawless, McCann, McIndoe-Calder and others (2012) also note that this is mainly due to the indigenous, employment-intensive nature of SMEs. SMEs are of major importance to many economies; in Ireland SMEs account for 68 percent of employment and 99 percent of firms in the private sector (Lawless, McCann and others, 2012). Since the financial crisis, SMEs have been hampered by restrictive policies and regulations, in their efforts to secure credit. However, it has been recognised internationally that credit constrained businesses engage with less economically valuable and growth-enhancing activity such as job and employment creation than similar unconstrained businesses (Campello et al., 2010). This means that financial institutions sit at the heart of the economy in making decisions about what businesses to give credit to. They will do this first to maximise their own profits but also promote growth leading to a sustainable employment sector and economy. This requires financial institutions’ credit risk models take into account at all times the current economic factors that are influencing good and bad credit decisions.

2.5 Macro Area Features Affecting SME and Economic Credit Risk

Hackbarth et al. (2006) note that even though there is a substantial amount of literature focussed on understanding and developing credit risk, there is a dearth of research into how macro-economic factors affect credit risk. Hackbarth et al. (2006) find it strange based on anecdotal suggestions from institutions that the economic business cycle is an important feature when calculating the probability of a customer defaulting. One example of this is that during a recession, consumers are less likely to spend money on discretionary goods or luxuries and as a result the credit risk of businesses in this sector will most likely rise due reduced demand from consumers. In Hackbarth et al. (2006) they found that macro-economic conditions clearly have an impact on credit risk.

Fama (1986) notes that over the counter derivatives broker-dealers measure risk using individual counter-parties’ details but also make use of geographic data and performance indicators of other industry groups. The Derivatives Policy Group (1995) also

---

5 The Derivatives Policy Group was made up of representatives of CS First Boston, Goldman Sachs, Morgan Stanley, Merrill Lynch, Salomon Brothers, and Lehman Brothers
made recommendations that credit exposure was measured by geography and industry exposure when calculating credit risk.

Financial institutions can also suffer from a risk known as the *winners curse.* In banking, this is a scenario where other financial institutions’ credit risk models score a borrower too risky and will not lend to that applicant. That same borrower then arrives at “our bank” where “our model” scores them as a good or likely to repay borrower and decide to lend to them. As a result “our bank” will take on an excessive amount of risk that where other banks’ expected losses (Duffie and Singleton, 2012). Duffie and Singleton (2012) also discuss how banks can mitigate their risk to the winners curse by including metrics such as borrowing rates and credit risk concentration limits by area or location.

Based on the Italian market, (Di Pietro and Lusignani, n.d.) investigated if SMEs experienced different levels of default rates based on the region in which they were located. They investigated the business cycles of these areas and looked at identifying which macro-economic features were the most influential in predicting default. For their experiment, they divided Italy into five areas, centre, north-east, north-west, south, and the islands. In their analysis, they confirmed that there was statistically a significant difference between default rates in different areas. This can be illustrated in Fig. 2.2

![Graph](image)

**Figure 2.2:** Evolution of the Italian SME Default Rate by Area (1985-2005)
(Source: Di Pietro and Lusignani, n.d.)

From Fig. 2.2 one can see that the average default rate in the south of the country (green trend line) and islands (blue trend line) is significantly higher than in the north-west(red trend line) and north-east(orange trend line). They use the Kruskal-Walls test
to confirm that the differences between the default rates are statistically significant.

In Antunes et al. (2005), using Portuguese data, they found that macro-economic features such as the employment rate, short term interest rate and gross domestic product were useful when included in the predictive models to estimate the probability of default. Using these features allowed them to develop stress tests where they could run macro-economic scenarios that would have a negative affect on the economy. This is the same method that is widely adopted by the International Monetary Fund (IMF) in their Financial Stability Assessment Program (FSAP) which is used to assess a country’s financial sector resilience and capacity to manage financial crisis. They produce tailored recommendations of a micro and macro nature based on each country’s circumstances (Marston, 2001)

Ardic et al. (2011) discusses how competition in developed countries and instability in developing countries provide some of the biggest challenges to modelling credit risk for SMEs. This view of developing counties is supported by research carried out by Rocha et al. (2011) which provides evidence from financial institutions in North Africa and the Middle East, detailing the immaturity of their financial systems and lack of SME transparency as some of the main obstacles.

### 2.6 Data Mining and Predictive Modelling

Knowledge discovery is defined by Frawley et al. (1992) as the “as the non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data”. Knowledge discovery can be thought of as extracting some piece of insight or value from data that could not have been done using a simple query. Fayyad et al. (1996) outlined an approach called the Knowledge Discovery in Databases (KDD) which centres on extracting these useful patterns from data stored in large databases.
The goal at the end of the KDD process is to realise some value or extract some piece of insight. This is typically done through the data mining step of the process, but Fig. 2.3 above illustrates it is only one step in the overall framework of KDD. Steps such as data selection, pre-processing, transformation or data modelling must be completed prior to the data mining step.

There are methodologies and frameworks for data mining. One of these is the Cross Industry Standard Process for Data Mining, usually referred to by CRISP-DM (Shearer, 2000). This data mining framework is the one most commonly adopted by data miners to work out a problem. Some polls show it is the leading methodology used by data miners\(^6\). It can be seen in Fig. 2.4, that the CRISP-DM process is split into six main

steps or tasks; *business understanding, data understanding, data preparation, modelling, evaluation* and *deployment*. The next most popular framework is known as Sample, Explore, Modify, Model and Assess more commonly known as SEMMA (Azevedo, 2008). This solution has been developed by Statistical Analysis System (SAS) institute but is seen more as a list of sequential steps that can be used to build out data mining solutions. The big advantage with the CRISP-DM methodology over SEMMA (illustrated in Fig. 2.4), is that it does not restrict one from moving between the different steps, and the arrow wrapping around the process suggests that even after deployment the process can continue.

Since they were established over 20 years ago, the frameworks of KDD and CRISP-DM outlining steps to extract insights have grown and developed. There are a growing number of communities that continuously overlap. This can be illustrated below in Fig. 2.5 where one can see that data mining, statistics, artificial intelligence and machine learning communities all share some common values. As a result, one can consider data mining as a combination of KDD, machine learning, statistics and pattern recognition that may or may not leverage on databases. This results in the field of data mining being largely made up of data scientists, data analysts, computer scientists and statisticians (Coenen, 2011).

For this research, the primary focus is on a subset of the data mining process called predictive modelling. This is represented in Fig. 2.4. Predictive modelling centres on predicting future events based on past or historic data. The predictions are

---

trained using past real world events and are tested and evaluated on unseen real world data to see how they perform.

Predictive modelling has many applications across a wide range of domains: some examples include election outcomes ((Silver, 2012); Tumasjan et al. (2010)), predicting how oil slicks spread (Liu et al., 2011), cancer prediction (Delen et al., 2005), and recently predictive modelling has become popular in sports predictions of baseball (Lewis, 2004), basketball (Stekler et al., 2012) and horse racing (Silverman and Suchard, 2013).

Financial institution commonly utilise predictive modelling across a variety of domains e.g. marketing and risk. In risk areas, banks will build predictive models to predict default probabilities for existing customers and evaluate customers applying for lending. This can reduce their risk, allowing them to increase profits meaning they can offer customers a better service and more personalised products. Thus credit scoring is one of the most used application fields for data mining (Baesens et al., 2009).

Predictive models are built by using interval/numerical or/and categorical/nominal features/variables/predictors/attributes that can explain the target/class/outcome to be predicted. Once the model is trained on historic data, model performance and evaluation is carried out on unseen data for testing. The data used in training will not be the same as in test thus the model may not generalise well on unseen data. In the literature, this is known as model over-fitting on the training dataset. Methodologies, frameworks and testing can be put in place to mitigate the risk to this issue. These will discussed further throughout this chapter.

2.7 Dataset Construction

Experts often detail how steps performed in the data construction and preparation stage can be some of the most time consuming when building predictive models. There are many steps that need to be considered during the data preparation stage for building a prediction model that models credit risk of SMEs. Two that will be discussed in this section are the sampling period and class label definition.

2.7.1 Sampling Period

As previously stated, predictive models are built using historical data. It must be acknowledged that past performance can be a useful predictor of defaulting but it does not guarantee that future predictions of the model will be accurate or reliable. A training dataset is built to train a predictive model where customers are observed at two different points in time (Martens et al., 2010). They are observed at the time that the prediction is made based on past performance. They are then observed sometime in the future called the default observation point at which they will be classified by the model as good or bad. The amount of time between these two points is commonly known as the outcome window. The length of the outcome window can vary based on business objectives and requirements. The industry standard in AIB dictates that this is usually 12 months.

2.7.2 Class Label Definition

Defining a customer as defaulted is dependant on the objective of the particular predictive model and the requirements of the financial institution (McNab and Wynn, 2000). The Basel II definition (paragraph 452) which is widely used by financial institutions, including AIB considers a default to have taken place when either or both of the following criteria are met:

- The bank/financial institution considers that the obligor is unlikely to pay its credit obligations to the banking group in full, without recourse by the bank to actions such as realising security (if held).

- The obligor is past due more than 90 days on any material credit obligation to the banking group. Overdrafts will be considered as being past due once the customer has breached an advised limit or been advised of a limit smaller than current outstandings.

According to Anderson (2007), there are two well known approaches to class label definition that financial institutions can choose: (i) a current status label definition which classifies a customer to have defaulted or not at the end of the outcome window or (ii) a worst status label definition which classifies whether the customer has defaulted or not throughout the outcome window. It is AIB’s industry standard to use the worst status
option. This is in keeping with Basel II (BCBS, 2005), that customers’ 90 days worst status covering a one-year period is considered the standard definition for customers who have defaulted.

2.7.3 Segmentation

Segmentation is carried out by splitting the dataset population into multiple groups and building a prediction model for each group (Myatt, 2007). This splitting will be carried out using group specific criteria and allows for modelling characteristics and features that are important to each group independently.

For marketing, Wedel and Kamakura (2012) illustrate how segmentation can be utilised to collate customers into homogeneous groups based on those customers’ buying pattens and demographic data like location, age and income.

Segmentation can be used in credit scoring to allow the lender to have more flexibility when personalising credit products for customers, for example interest rate and repayment structure (Kennedy, 2013).

Segmentation is carried out by analysts who generally require both subject matter expert experience and by leveraging statistical methods (Siddiqi, 2012).

2.8 Predictive Models

This section details some of the classification algorithms that can be used when modelling a binary classification problem. For this thesis the problem is if a customer defaulted or did not default on their financial obligation; hence, it is a binary classification problem.

The algorithms discussed in this section are not an exhaustive list but all are suitable for use in the financial industry. Classification models discussed include linear and logistic regression, k-nearest neighbour (KNN), support vector machines (SVM) and neural networks.

Based on research and experience working in industry, logistic regression is one of the most widely used classification algorithms used by industry.
2.8.1 Regression

Regression models are used to model the linear relationship between features in a feature space or between the features and the target variable.

A very simple form of linear regression is where there is one independent and one dependent variable, which is the target one is attempting to predict. The model is often represented by the following equation

\[
\text{Linear Model} = y = b_0 + b_1 x
\]

where the model is trying to predict \( y \) using the value of \( x \).

Fig. 2.6 illustrates a very simple real life example of linear regression. It demonstrates the linear relationship between a person’s heights and a person’s weight.

![Simple Linear Regression](image)

**Figure 2.6: Simple Linear Regression**

It can be seen from the above that \( y \) in this example is weight (wgt) and \( b_0 + b_1 x \) is \( 80 + 2 \times \text{Height}(hgt) \) is what we are using to predict \( y \). This is a very basic example but demonstrates how this can be leveraged for more complex feature sets. Predicting arrears is a dichotomous problem meaning the outcome of the experiment can only have two possible values.
2.8.2 Logistic Regression

Logistic Regression (See: Hosmer and Lemeshow, 2000) within the credit scoring industry is one of the most used algorithms (Hand and Zhou, 2010). As seen above in Fig. 2.6, a simple regression model outputs a continuous response, in the given example it is body weight. Credit scoring or predicting arrears is a problem where there can only be two possibly values default or not-default. To simplify, this is reduced to a binary problem where the outcome will be 1 or 0 (Zou, 2004). To transform the output of a regression model from $[-\infty, +\infty]$ to a probability between 0 and 1, a logistic transformation is applied. The logistic function can be used to take any value between $+\infty$ and $-\infty$ and output a value between 0 and 1. Fig. 2.7 below illustrates what a logistic function looks like.

![Figure 2.7: Standard Logistic Regression](image)

The logistic function is defined in Equation 2.2 as the following:

$$\text{Logistic Model} = p = \frac{1}{1 + e^{-(b_0 + b_1 x)}}$$  \hspace{1cm} (2.2)

As discussed previously, linear regression is an unsuitable classifier for making dichotomous predictions as linear regression produces predictions for a range beyond 0 to 1. Logistic regression also produces a curved line that is bounded by values between 0 and 1 (See Fig. 2.8).
In Fig. 2.8 the constant, \( b_0 \), dictates the position of the curve which can be moved left and right depending on its value, \( b_1 \) will be the slope of the curve.

The logistic regression model can be extended to include any number of interval and nominal features. This is illustrated in Equation 2.3.

\[
p = \frac{1}{1 + e^{-(b_0 + b_1 x_1 + b_2 x_2 + \cdots + b_p x_p)}}
\]  

Logistic Regression can also be used in cases where there are more than two outcome groups. For example, it could be used in predicting at what stage a customer is in the customer lifecycle e.g. Awareness, Interest/Consideration, Evaluation/Purchase. This is referred to as multinomial logistic regression.

One of the main major attractions of logistic regressions is that it allows one to use discrete, continuous features or a combination of both (Lee, 2005).

### 2.8.3 K-Nearest Neighbour

The \textit{k-nearest neighbour}, or k-NN for short, is an algorithm that classifies observations based on how its nearest neighbours are classified. It can be known as the nearest neighbour but in the majority of cases it is useful to use more than one neighbour (Henley and Hand, 1996). The intuition behind this algorithm is that instances which are close by each other will more likely be classified the same way (Cover and Hart, 1967). We can see in Fig. 2.9
Chapter 2. State-of-the-art

that results of the algorithm can vary with the choice of $k$. It should also be noted that when applying this algorithm to a binary experiment, it is good practice to choose only odd values of $k$, as this will eliminate the risk of ties from the decision process (Keller et al., 1985)

While adopting the k-NN algorithm there are multiple methods to decide what are your nearest neighbours. There are some common distance measures for continuous features only. Equation 2.4 is the Minkowski distance, which is one of the most common, where $p=1$ this becomes the Equation 2.5, the Manhatten distance, and where $p=2$ this becomes the Equation 2.6 the Euclidean distance.

\[
\text{Minkowski Distance} = \left( \sum_{i=1}^{k} (|x_i - y_i|)^p \right)^{\frac{1}{p}}
\]  
(2.4)

\[
\text{Manhatten Distance} = \sum_{i=1}^{k} |x_i - y_i|
\]  
(2.5)

\[
\text{Euclidean Distance} = \sqrt{\sum_{i=1}^{k} (x_i - y_i)^2}
\]  
(2.6)

The results from k-NN will vary depending on your choice of distance measurement. There are also other distance metrics for continuous features such as Correlation Similarity and Cosine Similarity (Sarwar et al., 2001)

Through analysis one can evaluate the optimal value for $k$, one based on inspecting the results and creating benchmarks. Anecdotally, the larger the value of $k$ the more precise
the algorithm can be but as with most things in data mining there are no guarantees. Choosing a high $k$ could also potentially cause over-fitting of the training data.

### 2.8.4 Decision Trees

The *decision tree* algorithm classifies observations into classes in the form of tree-like structure, hence the name. The algorithm seeks to partition the dataset into smaller subsets, using the relationship between the feature set and target variable to do so. An example of a simple decision can be seen in Fig. 2.10

![Simple Decision Tree for Yes, No Prediction](Source: Quinlan, 1986)

The output of the algorithm splits the data into smaller subsets of data; the output is a tree with a root node, internal nodes and leaf nodes. As can be seen in Fig. 2.11, the root node in this example ‘Outlook’ is the first node in the tree which means it is the most predictive feature.

![Decision Tree with Nodes and Leaves labelled](Fig 2.11: Decision Tree with Nodes and Leaves labelled)
The root node will have two or more branches. In this case, there are three ‘Rainy’, ‘Overcast’ and ‘Sunny’. Below these branches there are internal or split nodes ‘Windy’ and ‘Humidity’ which in turn output more branches. The bottom nodes of each decision or branch is the prediction or classification, this is called the leaf node. In this example the leaf node decision will be whether or not it will rain represented by Yes/No in Fig. 2.11.

The algorithm that builds decision trees is called the iterative dichotomiser 3 or more commonly known as ID3 (Quinlan, 1986). The algorithm applies a top down approach to choosing its root and internal nodes. The algorithm only evaluates one step ahead from where it is in the decision process at any time and does not allow for any backtracking. This decision making process is known as a greedy approach as it just makes the optimal solution at that particular stage of the process. Due to these limitations, the optimal solution is not guaranteed (Friedman et al., 1996).

The ID3 algorithm works by calculating the entropy and information gain at each step or decision node, where one uses the feature with the smallest entropy or feature that maximises information gain.

Entropy $H(S)$ seen in Equation 2.7 measures how much uncertainty there is in the data (Shannon, 2001)

$$H(S) = - \sum_{x \in X} p(x) \log_2 p(x)$$  \hspace{1cm} (2.7)

Where:

$S$: The current dataset for which entropy is being calculated, this will change each time entropy is calculated

$X$: The set of classes in $S$

$p(x)$: Proportion of observations in class $x$ compared to total number in set $S$

If $H(S) = 0$ then the observations in $S$ are all of the same class. Entropy is calculated for each feature and the feature with the smallest entropy is used to split at that step.

Information gain is used to measure the decrease in entropy after the dataset is split on a feature. The equation for the information is
\[ IG = H(S) - H(T) \] (2.8)

Where:

- \( H(S) \) is the entropy of \( S \)
- \( H(T) \) is the entropy of subset \( T \) based on splitting data on some feature

Information gain is calculated and the feature with the highest information gain is chosen to split the dataset. The algorithm then runs recursively until all the data is classified and predictions have been made.

Multiple decision trees may output the same results. This can be seen in Fig. 2.12 where two different decision trees classify the dataset correctly. Quinlan (1986) suggests that in scenarios like this, the simpler decision tree would be chosen (Fig. 2.12a).

![Simple Decision Tree](image)

![Complex Decision Tree](image)

**Figure 2.12: Simple and Complex Decision Tree Comparison**
(Source: Quinlan, 1986)

This is done because the simpler the rules of the tree the more likely the tree is to generalise well on unseen data. In other words, if the tree is too complex it is more than likely over-fitting the training dataset. There is also a computational cost to classifying complex algorithms.

Other issues one needs to be mindful of when building a classifier using a decision tree include the following. Information gain can be biased to features that have a large number of values. These features will result in a root node that produces a very broad or wide tree which classifies the training data well or perfectly but performs very poorly
on unseen cases. One scenario where this could happen would be if one used the unique identifier of each record as a training input; this model would perform very well in training but would not perform well on unseen data. There are methods to mitigate the risk of over-fitting against these features such as gain ratio, symmetric uncertainty and the Gini index. Quinlan (1986) noted that using these methods for node decision often produced favourable results when compared with information gain. It should also be noted that gain ratio, symmetric uncertainty, Gini index, and information gain can be used in the feature selection process which is discussed in Section 2.9.

2.8.5 Artificial Neural Networks

The artificial neural network (ANN) is a learning algorithm based on an understanding of how neural networks, such as our brains, learn. Motivation to study how an ANN works comes from the success of how the human brain is faster than the world's fastest computers at certain applications such as object recognition, speech recognition and general perception (Haykin, 1998).

As a human grows, their brain develops, learns and creates a set of rules based on the experiences it has had. These rules and experiences are stored in approximately 100 billion neurons or nerve cells in the brain. These neurons are connected in a network and use this as a method of communication, sending electrical and chemical signals back and forth between each other. On their own, each neuron is not very useful but in combination with other neurons and communication, this has allowed humans to learn and grow so successfully.

An ANN seeks to replicate the neural network of the human brain, albeit on a much smaller scale. It does this by taking advantage of powerful computers which carry out a lot of simple tasks very quickly. ANNs have proven their value from their ability to map out any non-linear function (White, 1989) and their prowess in applications such as pattern and speech recognition and forecasting (Kaastra and Boyd, 1995).

Fig. 2.13 shows a common layout of an ANN. As in the human brain, the ANN is comprised of processing neurons usually known as nodes which are organised into three layers, input, hidden and output. Nodes are connected between layers and as seen in Fig. 2.13 each connection may carry a different weight.
Data comes in through the input layer’s nodes and is fed through the network, from the input to the hidden and then onto the output layer. In the hidden layer, each node calculates a sum based on the input node and the weight of the connection. These hidden nodes then pass on values to the nodes in the outer layer where another calculation is performed: this calculation converts the value to a value between 0 and 1 by passing it through the sigmoid function seen in Fig. 2.7. Throughout the training process, the connection weights are changed and tested in order for the ANN to learn and improve its predictions (Haykin, 1998).

ANNs have become increasingly popular in recent years due to improvements in the algorithms, the increase in computer power and success in application such as object and speech recognition. However, some remain sceptical because of the “black box” nature of their results, where users do not know what the internal workings of the algorithm are (Kaastra and Boyd, 1995).

2.8.6 Support Vector Machines

A Support Vector Machine (SVM) algorithm was developed first by Vapnik (1995). The algorithm performs classifications via a hyperplane in a higher dimensional feature space that maximises the margin or distance separating the two classes. It can be seen in Fig. 2.14 that the two classes are linearly separable using the hyperplane in the higher dimensional feature space.
SVM handles situations of non-linear data by using kernel functions (non-linear) to transform the data into a higher dimensional feature space. This allows it to become linearly separable via a hyperplane, this is known as the \textit{kernel trick}. This is illustrated in Fig. 2.15, where non-linearly separable data is transformed into a higher dimensional feature space where it can be linearly separated using a hyperplane. This ability is what differentiates SVM from logistic regression.

There is much documentation illustrating the successful application of SVMs in several domains, including areas such as credit risk evaluation (Van Gestel and Baesens, 2009) and text categorisation, cancer diagnosis and pattern recognition (Shin et al., 2005).

\subsection*{2.8.7 Ensemble Models and Boosting}

In 1907, mathematician Sir Francis Galton went to a market in which there was a challenge to approximate the weight of an ox. After evaluating the 787 forecasts made
by the participants, he noticed that while there was a large variance in the forecast from the correct weight, the median value of the forecasts was less than 1% away from the correct weight of the Ox (Galton, 1907). Although separate forecasts failed miserably, the united wisdom of the every guess generated a very accurate estimate. This is similar to how ensemble models are generated.

Boosting relates to a powerful principal of generating a very accurate prediction model using an aggregation of reasonably inaccurate “rules of thumb” (Freund et al., 1999). A frequently utilised ensemble is the Adaptive Boosting algorithm which is frequently called AdaBoost in the literature. A weak learner is produced from the first iteration where all of the observations can potentially be chosen. For later sampling of the distribution, the model makes adjustments based on the error rates of the classifier, so that model only will look at samples of data that were incorrectly classified (Freund et al., 1999). In this way, the algorithm is adopting step by step, based on the errors of past classifications and then focuses on correcting the record labels it got incorrect. A specific method of AdaBoost is the Boosted-Stumps Model. This is an ensemble model that leverages decision tree stumps, which are decision trees with one split. This method is seen to be optimal compared to the common AdaBoost model which tends to over-fit the training dataset (Caruana and Niculescu-Mizil, 2006).

Bagging is an alternative method employed to generate ensemble models leveraging decision trees. One method of bagging is bootstrap replica. This method works on the principle for a dataset of size \( n \) to be used to train a model, generating trees using just different partitions of the training data with replacement (Dietterich, 2000).

Random forests are another example of an ensemble model generated using multiple decision trees. A large number of decisions trees will be trained and results are combined together to make a classification, hence why it is called a ‘forest’. Each decision tree is trained on random subsets of the features available, hence why its called a ‘random’ forest. Research provided by Breiman (2001) demonstrated that the random forests performed better on tests compared to the AdaBoost method across a variety of datasets.
2.9 Feature Selection

Feature selection is a process of choosing the best subset of features from the full dataset to train the prediction model. Guyon and Elisseeff (2003) discusses that feature selection methods are usually split into one of three categories (i) filter techniques (ii) wrapper techniques and (iii) embedded techniques.

Filter Methods

Filter feature selection methods use statistics to assign a score for each feature versus the target. The features are then ordered by predictability and a decision is made as to what features to keep. Filter method techniques include information gain, correlation coefficient and the chi squared test.

Wrapper Methods

Wrapper feature selection methods evaluate various subsets of features together while scoring the model for each subset. The resultant different model results are then evaluated and compared against the other results, returning the result which offered the best score based on the model evaluation criteria. Forward, backward and general stepwise regression are very common techniques of wrapper methods.

Embedded Methods

Embedded feature selection methods attempt to combine the two previous methods. That is, the method looks to learn what features are useful as the model is being created.

Feature selection is very important in the credit scoring process and there are many reasons in the literature that suggest it should be used. The curse of dimensionality is one such issue. If there are too many features in the model, it may perform well on the training dataset but when executed on unseen data it may perform poorly because the model has over-fitted many irrelevant or noisy features (Loughrey and Cunningham, 2005). Research from (Thomas (2009); Mays (2004)) advises that to build a robust scoring model there should be somewhere between 8 and 20 predictive features. There
are many reason for this: there is a practical issue of having to model more features; there are costs, overheads and maintenance associated with each redundant feature in the model. Referencing Fig. 2.12, one can see that it is much more desirable to have a simple decision tree than a complex decision tree with the same outputs. Similarly, if two datasets are providing the same results and one is a subset, it is always better to choose the subset.

2.10 Coarse Classification

Coarse classification is often utilised to transform the predictive features into a simpler form which is better suited for modelling (Carroll and Ruppert, 1988). Coarse classification is also referred to in the literature as binning, grouping or discretisation. For continuous and categorical features, values are transformed into a small number of bins or groups. These values are mapped into these groups by referring to the target feature to identify the optimal cut-off points.

Coarse classification has many benefits, one of which is that it allows for capturing the features’ non-liner relationship with the target feature. This is achieved by each category in the group being treated as its own dummy variable which will have its own weight in a logistic regression model (Hand et al., 2005). Coarse selection can also increase the overall robustness of the model by reducing the possibility of over-fitting. It does this by creating groups with the optimal number of good records (Baesens et al., 2009). It also offers the capability of mitigating against the risk of outliers and missing values.

Coarse classification is also very quick to deploy. Previously, it would have been achieved by analysts iteratively completing the process which was a very time consuming. Algorithms now such as chimerge and recursive partitioning to name a few, find optimal cut-points in the features quickly and accurately. These groups are then evaluated using the Weight of Evidence (WoE) and Information Value (IV) (Garcia et al., 2013).

A common methodology in coarse classification is to break up each feature into roughly three to six bins or groups (Hand et al., 2005). It is recommended that bins are limited to six to ensure the model does not become over-parameterised and difficult to manage. Conversely, the model will become too rigid if fewer than three groups are used.
Approximating the WoE of each bin and tuning where optimal is the most-known method in carrying out course selection (Thomas, 2009). In a credit scoring problem the WoE for bin \( i \) is defined as

\[
\text{WoE} = \ln \left( \frac{n_g(i)}{n_b(i)} \right) / \left( \frac{N_g}{N_b} \right)
\]

(2.9)

where the amount of goods in a bin \( i \) is \( n_g(i) \), and the amount of bads in a bin \( i \) is \( n_b(i) \). \( N_g \) and \( N_b \) are the total amount of goods and bads in the full dataset. A negative WoE signals that a bin is more likely to default, whereas a positive WoE signals they are likely to not default.

The information value is frequently applied with the WoE. The IV is a powerful method for ranking variables by their importance which can be used for feature selection in the prediction model. The IV suggests the predictive capability of a binned feature and is defined as

\[
\text{IV} = \sum_{i=1}^{j} \left( \frac{n_g(i)}{N_g} - \frac{n_b(i)}{N_b} \right) \times \text{WoE}_i
\]

(2.10)

where \( j \) is the number of bins in a feature. The IV of each bin is known as a contribution, which are then added together to generate the IV of a feature.

In general, binned features with an IV between 0.3 and 0.5 are thought to be very predictive features (Mays, 2004). If the IV is more than 0.5 it could be an anachronistic feature and should be examined further (Siddiqi, 2012). Binned features with an IV smaller than 0.1 are thought to be weaker features and their removal from the model should be considered (Anderson, 2007).

### 2.11 Class Imbalance Problem

A key assumption which needs to be taken into account when using classification algorithms is that there is a balanced distribution of the target class (Japkowicz, 2000). Target class imbalance is described by (Chawla et al., 2002) where the number of records
in each class are not equal. In a balanced dataset, the ratio between the a binary target
class would be close to 50:50.

One of the issues with imbalance arises when algorithms assume there is a balanced target
class and they attempt to maximise the accuracy by predicting the most common class
(Drummond and Holte, 2005). The algorithms attempt to minimise the classification
errors but fails to account for the incorrectly classifying cases (Seiffert et al., 2009).
While the overall classification might be very accurate, the results are not very useful
in real world problems. This is because, in most cases, the algorithm will focus on the
majority class, because of how heavily it is weighted in the training dataset and therefore
ignoring the minority class. This is a serious issue because in most situations, the aim
will be to predict the minority class. In this thesis, SME customers going into default is
the minority e.g. there are more SME customers who do not default at the end of the
outcome window than SME customers who default.

In recent years, more of the data mining and machine learning literature has explored
the issue of class imbalance. Weiss (2004) discusses the role and issues that rare class
instances can play in data mining. Weiss (2004) makes the distinction that there are
two types of class imbalance which depend on the type of rarity in the data, these are
called absolute rarity and relative rarity.

The primary issue with rarity is that there is simply a lack of data in real world problems.
Absolute rarity occurs when the number of instances related to the rare class is very
small in an absolute sense. Lack of data means it is difficult to identify what leads to a
rare class. Fig. 2.16 illustrates how absolute rarity can pose difficulty.

![Figure 2.16: Impact of Absolute Rarity in Data Mining](Source: Weiss, 2004)
On the left side of Fig. 2.16 there is only one rare/positive example, compared to the right where there is more data thus more rare cases. It can be observed that the decision boundary on the right side of Fig. 2.16 is much more accurate when there is more data than on the left side Fig. 2.16 when there is just one observed rare class. This is a simple illustration of the fact that more data should facilitate better predictions. Relative rarity is where classes are not rare in an absolute sense but are rare relative to other objects. A supermarket example can illustrate this better: imagine trying to identify the relationship between two items but these items are rarely purchased as a whole, so even if they happen to be purchased together, the relationship may be difficult to identify.

As previously mentioned, class imbalances in the dataset occur very often in real world problems and thus research has been devoted to proposing methods of mitigating against this risk. Chawla et al. (2004) proposed solutions centres on fine tuning the algorithm and manipulating the data.

**Manipulating the data**

A method of manipulating the data is to resample the data with the aim of balancing the distribution of the target class. Solutions proposed are as follows:

- Random undersampling of the majority class
- Random oversampling of the minority class
- Synthetic sampling of the minority class

A method commonly used is to randomly oversample the minority class in the training set. However, this increases the chances of over-fitting the algorithm to the training data as the model has been trained on multiple copies of the same data, none of which are adding any new information. This may cause the trained model to be biased and skewed on the training data, causing it to perform poorly on the test data (See Hawkins, 2004).

Random undersampling of the minority class is where random samples of the training dataset that are part of the majority class are removed. This means the number of
minority classes remains unchanged but the majority class is reduced, therefore the overall target class will become more balanced. The issue that arises from undersampling is that there is a possibility that important information from the training dataset will be removed. Kennedy (2013) details that undersampling the majority class is not a useful solution for the issue of absolute rarity.

Synthetic sampling is an alternative method to randomly oversampling the minority class. New data items are added to the training dataset but unlike oversampling, which adds duplicate records, the records added are dummy or made up in a way to look similar and take characteristics of the already existing records, thus they are not duplicates but synthetic. One method for creating synthetic data was proposed by (Chawla et al., 2002) where data was generated by creating data items using \textit{k-nearest neighbours} (KNN) where the item would sit between minority classes.

![Figure 2.17: Example of the K-NN for $x_i$ using $k = 6$. Data created using SMOTE based on the Euclidean distance.](Source: He and Garcia, 2009)

Above in Fig. 2.17 above illustrates how synthetic data using the SMOTE methodology can be generated.

**Fine tuning the algorithm**

Some methods to handle the class imbalance issue do so by fine tuning the algorithm. One method which is illustrated in Chapter 4 of this thesis is to adjust the cut-off or threshold value for the model on which some performance measures are dependant. Provost (2000) warns that it would be \textit{“critical mistake”} not to do so if faced with this issue. Chawla et al. (2004) suggests using evaluation measures (such as accuracy) that rely on a specific threshold, could lead to misleading results when the target class is
imbalanced. They instead recommend using ROC (Receiver Operating Characteristic) and AUC (Area Under Curve) to get more accurate and conclusive predictions.

2.12 Model Validation Methods

In data mining, historic data is used to train a model to make future predictions. Trained classification algorithms like the ones already discussed in this chapter need to be validated and tested. This section details some of the methods and approaches to tackling this problem, such as those discussed extensively in (Refaelzadeh et al., 2009).

**Holdout Validation**

The *holdout validation* method is used to split the dataset into partitions, one for training and one for testing. The algorithm is trained on the training partition. This allows for the model testing to be carried out: when the model classifies the test partition to evaluate how well it generalises and performs on unseen data. This method is not faultless however, particularly where not all the data is used for training and because results can be dependent on how and what datasets are used for training and testing. Examples where this could become an issue is if important information in the data for training is lost in the test partition, or the instances chosen for test may be too easy or too difficult to classify. These situations may cause results to be skewed and cause prediction bias in that testing partition.

The hold-out method can also incorporate a third subset called the *validation set*. In this case the dataset is partitioned into three partitions train, validation, and test. The validation set’s purpose is to fine tune model’s parameters, for example selecting the best threshold cut-off measure for a model or performance measure. It is not always required for building predictive models.
Fig. 2.18 illustrates how the complete dataset is partitioned into training and test datasets.

To address the biases in the holdout method, one can run tests multiple times with the results averaged, or more commonly employ a method such as *k-fold cross-validation* discussed in the next subsection.

**K-Fold Cross Validation**

The *k-fold cross validation* can be used to deal with the bias issues discussed above in the holdout method. The first step of this method is to split the data into $k$ equally sized partitions called folds. A model is then trained using $k$ iterations, where for each iteration a different fold of data is used for testing and training the model. This is illustrated below in Fig. 2.19 where there are 5-folds, and for each iteration you can see a different fold is being used for training and testing.
It is important and usually common practice that each fold is representative of the whole dataset, that is for each example that the target class ratio split is the same for each fold as for the entire dataset.

**Leave-One-Out Cross-Validation**

The *leave-one-out cross-validation* (LOOCV) method is a specialised version of *k*-fold cross-validation, where *k* is equal to the number of observations in the dataset. In layman’s terms, this means all the data except for one observation is utilised in training the model and testing is done on one observation. This is completed for each observation in the dataset. Although it is worth noting that accuracy estimated using this method produces unbiased results, it also outputs high variance which can lead to misleading results (Refaeilzadeh et al., 2009).

**Repeated K-Fold Cross Validation**

The *repeated k-fold cross validation* method is another specialised version of *k*-fold cross-validation. In an attempt to improve the performance of the model, in this method *k*-fold cross-validation is re-run multiples times. Each time it is re-ran the data is shuffled so data will appear in different folds for each repeat run.

**Summary**

The holdout and cross validation methods are both used extensively for evaluating the performance of the prediction models. If there is a large enough dataset, then the holdout method is usually a good choice for measuring model performance. When there is a small amount of data available it better to use *k*-fold cross validation. *K*-fold is theoretically simple to understand but may not be as simple to implement as the code may be tedious and time-consuming meaning the gains may not be worth the investment in industry.

Kohavi and others (1995) and Salzberg (1997) both evaluated approaches to choosing the best validation method. Kohavi and others (1995) analysed many cross validation methods, including regular, leave-one-out, stratified. They came to the conclusion that
10-fold cross-validation produced the most accurate and unbiased results. Salzberg (1997) also studied the issues of comparing model performance and proposes a solution that combines k-fold cross validation with appropriate hypothesis tests as opposed to evaluating the average accuracy.

### 2.13 Model Performance Measures

This section details some of the metrics used to assess the accuracy of a classifier. The result of the classification algorithm maps the modelled data into a category, in this thesis it is a binary classifier that has output of 1, identifying customers who will default (bad) or 0 signifying customers who will not default (good). The majority of classification algorithms will produce a ranked numeric value which can be converted to a binary representation by some threshold or cut-off, where the decision is driven from the business objective to be optimised. This section will begin with a description of a confusion matrix, detailing how it is leveraged to build other performance measures. It also explores how charts and metrics can be leveraged together to decide on the performance measure which maximises the intended objective.

#### Confusion Matrix

The results produced by the classification algorithm can be represented by a contingency table known as a confusion matrix. In this thesis, the classification algorithm will output a binary classification, so the confusion matrix will be a $2 \times 2$ matrix with two classes, known as the positive and negative class. For this study, the positive class will be the customers who default and the negative class will be the customers who do not default. The confusion matrix will illustrate what proportion of correct and incorrect predictions were made with respect to the target. The confusion matrix in this research can be broken down into the following categories:

- **true positive** (TP), cases that are predicted to default, and are correctly classified as positive
- **false positive** (FP), cases that are predicted to default, and are incorrectly classified as positive, also known as Type I error
Chapter 2. State-of-the-art

- **false negative** (FN), cases that are predicted to not default, and are incorrectly classified as negative, also known as Type II error
- **true negative** (TN), cases that are predicted to not default, and are correctly classified negative

Fig 2.20 illustrates how information from a confusion matrix can be presented and read.

![Confusion Matrix](image)

**Figure 2.20: Confusion Matrix**

Using the results from the confusion matrix, model evaluation measures can be calculated and evaluated for the required objective. Measures such as *accuracy* (Equation 2.15) which measures the proportion of predictions that were correct and the *misclassification rate* (Equation 2.17) which shows the proportion of predictions which were wrong. Other measures are *sensitivity* (Equation 2.11), otherwise known as the *recall* or the *true positive rate* (TPR), which measures the proportion of the positive instances that are correctly identified i.e. proportion of default cases that have been predicted correctly and *specificity* (Equation 2.12) which measures the proportion of negative cases that are predicted correctly i.e. proportion of non default cases that have been predicted correctly. Finally, there is *precision* (Equation 2.13) which measures what proportion are correct when the classifier predicts positive outcomes and *negative predictive value* (NPV) (Equation 2.14) that measures the proportion of negative predictions that were correct i.e. what proportion were correct if the classifier predicted the outcome would be non default. One measure that can be very useful when there is class imbalance is *balanced accuracy* (Equation 2.16), (Brodersen et al., 2010) discusses how using this negates the impact of bias or skewness from the more frequent class.

\[
\text{Sensitivity} = \text{Recall} = \frac{TP}{TP + FN} \quad (2.11)
\]

\[
\text{Specificity} = \frac{TN}{FP + TN} \quad (2.12)
\]
Chapter 2. *State-of-the-art*

Precision = \( \frac{TP}{TP + FP} \) \quad (2.13)

Negative Predictive Value = \( \frac{TN}{TN + FN} \) \quad (2.14)

Accuracy = \( \frac{TP + TN}{TP + FP + FN + TN} \) \quad (2.15)

Average Accuracy = Balanced Accuracy = \( \frac{\text{Sensitivity} + \text{Specificity}}{2} \) \quad (2.16)

Misclassification Rate = \( \frac{FP + FN}{TP + FP + FN + TN} = 1 - \text{Accuracy} \) \quad (2.17)

Fig. 2.21: Confusion Matrix Example

Fig. 2.21 illustrates how the TP, FP, FN, TN can be used to create performance metrics for a classification algorithm. As discussed above, confusion matrix based performance measures are built on the threshold that is selected for converting the predicted numeric scores into a binary outcome. Anecdotally, a cut-off or threshold of 0.50 might seem acceptable but this rule may not always apply, especially if there is an imbalance between the positive and negative class in the dataset. The cut-off should ideally be based on the business objective which will look to minimise, maximise and analyse the trade off as cut-off is altered.

A confusion matrix is not the only way to evaluate the performance of the classification algorithm and is not always advocated in the literature or by industry. In studies completed by Lessmann et al. (2008), a classification cut-off was not selected, arguing that studies comparing the same dataset and classifier could come to different conclusions.
As mentioned in this section the confusion matrix measures the performance of classifier at a specific threshold. This can be leveraged to create graphical representations of the overall fitness of the model at any threshold. One such method that will be discussed in the next section is the receiver operating characteristic (ROC).

As a confusion matrix is used to evaluate the model where output is divided into distinct or discrete categories, it can be reconstructed for any point on the ROC curve. ROC is also directly related to two other performance methods, area under the roc curve (AUC/AUROC) and Gini which will also be discussed.

**ROC Chart, AUC and Gini Coefficient**

The ROC chart is used to evaluate and illustrate how well the model fits training and test data. This can be a quick test to see if the model generalises well on test data. Fig. 2.22 illustrates this: on the x-axis the false positive rate and on the y-axis the true positive rate is represented. The points on the ROC chart are generated from the confusion matrix built from many cut-off or threshold value between $\theta \in [0, 1]$. Fig. 2.22 illustrates how the ROC chart is generated for varying thresholds.

![Figure 2.22: ROC Chart Example with thresholds 0.65 & 0.50](image)

Fig. 2.22 also illustrates how point (0,1) represents perfect classification, that is the classifier correctly predicting all outcomes.
The ROC chart is basically a combination of confusion matrices over many cut-off values of a classifier. As you can see above, a finite number of observations in the dataset dictates the number of thresholds which can be used to generate a ROC chart.

To compare ROC chart results of different classifiers, the area under the ROC curve (AUC) is utilised (Bradley (1997); Hanley and McNeil (1982)). In Fig. 2.22 the area under the blue ROC line represents the AUC, which is intuitively the area under the curve. In the case of perfect classification this value will be 1, for a random classifier the AUC would be 0.5.

It is worth noting that AUC does not give total probability of the classifier. Its usefulness is linked to it being combined with the ROC to evaluate a classifier across training, validation and test datasets. For example if the ROC curve shifts significantly or is not similar from training to test data, this is suggestive to possible over-fitting and the model does not generalise well. Also it is worth looking out for a drastic change in the AUC from training to validation/test, as this is another sign the classifier does not generalise well and may not be useful for predictions. These attributes make it a very strong measure for classifier selection. ROC also do not tend to cross over, therefore when comparing two classifiers the one for which AUC is higher will be the better classifier, independent of the threshold or cut-off.

Fig. 2.23 demonstrates how the performance measures for the confusion matrix vary for different thresholds, but again bearing in mind that there is just one measure for the AUC of the ROC chart for this model.

A metric that is commonly used for performance measurement in AIB and industry for credit scoring is the Gini coefficient and this is discussed in (Hand, 2005). Gini equates to twice the area in between the diagonal of a random classifier and the ROC curve. The equation for this can be seen below in Equation 2.18
As when using a threshold, there are limitations to using the AUC and Gini to measure classifier performance. Although extremely useful for analysing the performance of a wide range of thresholds, it is not as useful when trying to maximise the performance over a narrow range of thresholds.

One statistic that is commonly used in credit scoring in AIB industry and in the literature is the *kolmogorov-smirnov* (KS) statistic, this will be discussed in the next subsection.

**K-S Statistic**

The Kolmogorov-Smirnov (KS) statistic measures performance of prediction models. The measure is a single value between 0 and 1. The KS evaluates the maximum difference between the cumulative positive and negative distributions of the predicted positive and negative class (Seliya et al., 2009). This means it evaluates, at each threshold of the model, the cumulative percentage of positive cases identified and the cumulative count of negative cases identified.

**2.14 Conclusion**

This chapter has summarised the relevant literature for a number of topics central to this research project.

Credit scorecards are generally employed by financial institutions to assess the risk of existing or potential customers defaulting on their financial obligation, allowing the institutions to make informed decisions about future loses. Credit score models are built using data such as borrowers’ repayment performance information and demographic details. These scorecards give lenders a standardised, structured, easy to interpret and transparent result which is used for assessing customer likelihood to default on the financial obligation.
There is no worldwide definition for what a small and medium enterprise (SME) is. In Ireland, the definition for a SME is a registered business with fewer than 250 employees, an annual turnover not above €50 million, and/or an annual balance sheet total not above €43 million.

The interdependency between consumer confidence and a strong SME sector were highlighted. It has been acknowledged that there is a lack of research into how macro-economic features affect credit risk. However, one study demonstrated that the differences between default rates by region in Italy were shown to be statistically significant. It was also found that there have been examples when unemployment rate and default rates have useful when used in modelling credit risk.

Data mining and predictive modelling encapsulates a large number of subject areas such as feature selection, classification algorithms, model validation and model performance. Logistic regression is the most commonly used classification algorithm for credit scoring when predicting how likely a customer is to default on their financial obligation. Similarly, coarse classification is the most common method used in credit scoring to select and evaluate features to be included in the predictive model. This technique creates bins within each feature and evaluates these bins’ relationship with the target feature. It is widely used because of the accuracy of the results and its ability to create a robust feature set for prediction which is able to cater for missing values and outliers. Area under the curve (AUC) is the most common method of evaluating the performance of the credit scoring models and is generated by measuring the performance of the model over all possible thresholds.
Chapter 3

Data

3.1 Introduction

This chapter presents the data that will be used for the experiments conducted in this research. This chapter is compromised of two main sections.

The first section will outline from where the customer data for the experiment have been gathered and under what criteria they have been selected. As part of the experiment, a baseline predictive model will be built, using features that were used in a historic industry credit scorecard models in AIB. The performance measure to evaluate the experiment will also be evaluated using the baseline model.

The second section will outline what macro-economic features will be built as part of the experiment in this research. The aim of this research is to investigate the predictive power of macro-economic features by geographic regions such as electoral divisions and local authorities in Ireland. To do this, the addresses that are stored in AIB’s databases are queried against a search engine and string metric algorithm application to map customers’ addresses to global positioning system (GPS) coordinates. The macro-economic features will be sourced and created from internal data sources in AIB and from external open data sources. This section will also explain how these features have been created, what transformations or data wrangling was required so the features fit/map into an analytical base table (ABT) discussed in Section 2.7. Care was also taken throughout this experiment to ensure that anachronistic variables were not included in any of the
predictive models built as part of this research e.g. features that contain information about the outcome after the observation point.

3.2 Experiment Set-up

This section will detail the data that is in scope for prediction in this research. In Fig. 3.1 below, SME customers are selected at the observation point, June 2014. These customers are not in default at this point in time. Information prior to the observation point will be used for modelling to predict if a customer is likely to go into default, which is known as the performance window.

Data from each individual customer’s performance will be taken from data in the performance window, which will be combined with macro location-based data prior to the observation point. This data will be aggregated and structured into features for an ABT. The aim will be that these features will be able to distinguish what customers are likely to default on their repayment in the next 12 months.

![Figure 3.1: Experiment Performance Window and Outcome Window](image)

Section 2.7.2 described the two methods used to define if a customer was in default or not: (i) the worst status label definition method and (ii) the current status label method approaches. As mentioned previously, this experiment will be using the industry standard worst status method, which means if the customer is in 90+ days arrears at any stage in the outcome window they will be labelled as a bad customer or as one that has defaulted on their financial obligation.

3.3 Customers for Credit Scoring and Existing Features

The customer data used for prediction in these experiments was sourced from a financial institution, AIB which is one of the two main pillar banks in Ireland. It contains details
of 27,082 SME customers who were active between June 2014 and June 2015. These
27,082 customers are a subset of SME customers on the AIB book, as the experiment
will only be completed on one of the loan systems in the financial institution. The
customers’ default status at June 2015 will be recorded and used as the target for
building the prediction model. The baseline model for this experiment will be built
from features from a historic scorecard in AIB.

As mentioned in Section 2.7.3, it is very common in credit scoring to model the pop-
ulation into multiple groups. This is done to ensure that homogeneous customers are
grouped and modelled together based on such things as pattern, characteristic, demo-
graphic. This is common practice in industry. In AIB one method of modelling the
credit risk of customers is to split the customers into two segments. The criteria for
selecting in which segment each customer is modelled in is if that customer has been
in arrears previously or not. For for this research, a customer will be modelled in the
Previous Delinquency segment, if the customer has not been performing well and been
in arrears previously. If the customer has been performing well and not been in ar-
rears previously they will be modelled in the No Previous Delinquency segment. The
historic scorecards that were used in AIB utilised various features to build a model for
each segment population. This was done using empirical analysis of the data where it
was observed that different features contributed to the prediction of each subset of the
population with a few overlapping features in each subset. Due to sensitivity of the
information in AIB, the feature set for these models will not be documented in this
research paper.

As explored in Section 2.11, class imbalance in datasets is a major real world problem
when building a predictive model. This happens when the target class is not distributed
evenly in the dataset. The dataset in this experiment suffers from this imbalance.
However, because the data is partitioned into two segments, the imbalance in previous
delinquency dataset improves significantly but gets worse in the no previous delinquency
dataset.

The characteristics of the two datasets to be used to build the baseline benchmark are
illustrated in Table 3.1 below.
Chapter 3. *Data*

<table>
<thead>
<tr>
<th>Model</th>
<th># Numeric</th>
<th># Nominal</th>
<th># Observations</th>
<th># Good</th>
<th># Bad</th>
<th>Good:Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Delinquency</td>
<td>11</td>
<td>0</td>
<td>2,926</td>
<td>2,198</td>
<td>738</td>
<td>75:25</td>
</tr>
<tr>
<td>No Previous Delinquency</td>
<td>9</td>
<td>0</td>
<td>24,156</td>
<td>23,505</td>
<td>651</td>
<td>97:03</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>27,082</strong></td>
<td><strong>25,703</strong></td>
<td><strong>1,389</strong></td>
<td><strong>738</strong></td>
<td><strong>651</strong></td>
<td><strong>95:05</strong></td>
</tr>
</tbody>
</table>

Table 3.1: Characteristics of datasets to be used in the exploratory evaluation for training a benchmark model and assessing the evaluation metrics to be used in the research

# Numeric refers to the number of continuous features

# Nominal refers to the number of categorical features

It can be seen above in Table 3.1 that the two datasets are not very similar. There are only 2,936 customers in the *previous delinquency* dataset and 24,156 in the *no previous delinquency* dataset. Perhaps the biggest difference is the class imbalance difference between the datasets. 25% of customers in the *previous delinquency* dataset are bad by the end of the outcome window but there is only 3% of the *no previous delinquency* dataset that are bad at the end of the outcome window. As discussed in Section 2.11, this presents significant challenges when trying to build a predictive model but is also a very common challenge in real world applications.

### 3.4 Macro-Economic Areas for Experiment

The experiment in this research is to investigate if macro-economic features by locations are useful for predicting if SME customers are likely to go into default. The two macro-economic regions for features are *Electoral Division* (ED) and *Local Authority* (LD). Below Fig. 3.2 maps out the electoral divisions and local authorities in the Republic of Ireland.
There are 34 primary local authorities in the Republic of Ireland, including 29 county councils and 5 city councils. Organisations within each area are responsible for managing issues such as housing, planning, roads, water supply and sewerage, development incentives and controls, environmental protection, recreation facilities and amenities, agriculture, education, health and welfare\(^1\). There are 3,440 electoral divisions in the Republic of Ireland. Electoral divisions are formed by grouping town-lands together and are the smallest legally defined administrative areas in the state from which small population statistics are published in the Census\(^2\).

### 3.5 Converting Addresses to GPS Coordinates

A mechanism is required to link macro-economic features by electoral division and local authorities to a SME customer. In an ideal world this would be done through a data

\(^1\)http://www.iro.ie/local_authorities.html
\(^2\)http://census.cso.ie/censusasp/saps/boundaries/eds_bound.htm
model where a customer’s address would be linked to the address reference database. This can be illustrated through the Entity-Relationship Model (ERD) in Fig. 3.3.

Unfortunately, this is not the case currently in AIB, where addresses are stored in free text fields across multiple systems. This can be illustrated in Fig. 3.4 where one can see unstandardised and free text addresses are stored in the customer table instead of in an address reference table as in Fig. 3.3.
With the release of Ireland’s new postcode system in July 2015 (Eircode\(^3\)), AIB has looked to position itself strongly for its deployment and how it would integrate into AIB’s current legacy systems. For this reason an investment was made in the GeoDirectory database\(^4\). It is a product established by An Post\(^5\) and the Ordnance Survey Ireland\(^6\). It provides a complete database of all the addresses in the Republic of Ireland and geolocation details including 1.8 million buildings. The Eircode database is heavily connected with GeoDirectry database as it is essentially the same database with a new address identifier called Eircode. The database diagram for GeoDirectory is shown below in Fig. 3.5 which includes an Electoral Division and Local Authorities table.

\(^3\)http://www.eircode.ie/
\(^4\)https://www.geodirectory.ie/
\(^5\)http://www.anpost.ie/AnPost/
\(^6\)http://www.osi.ie/
To conduct this research a mechanism or application for matching customer addresses to the GeoDirectory database had to be built. There are many vendors in Ireland and internationally who provide services to correct and validate addresses such as Address Doctor\(^7\), Gamma\(^8\) and Data Ireland\(^9\). Committing to one of these products would require a project to evaluate each service where AIB would analyse the pros and cons, understanding the full requirements of the financial institution.

In the interim, as part of this research, it was decided to look at in-house solutions that could be developed using existing resources and open source technologies. After some investigation and experimentation it was identified that it was possible to build a solution leveraging an address database GeoDirectory, a search platform/engine Solr\(^10\) and a high level programming language Python \(^11\).

\(^7\)https://www.informatica.com/addressdoctor.html
\(^8\)http://www.gamma.ie/about-gamma
\(^9\)http://www.dataireland.ie/
\(^10\)http://lucene.apache.org/solr/
\(^11\)https://www.python.org/
Solr search works by creating an index of the data chosen for application. An example of this is illustrated in Fig. 3.6\(^\text{12}\).

For this experiment, the GeoDirectory database was indexed using Solr, allowing it to be queried through the web interface or multiple Application Program Interface (API) such as Python, JavaScript, Ruby, Java, HTTP. Solr returns a number of results from queries posted against it based on these indexes which can be seen in Fig. 3.7.

\(^{12}\text{http://blog.e-zest.net/about-apache-solr/}\)
Although this solution worked quite well in the majority of cases observed, it did have some issues because of the way Irish addresses are structured. For example, because there are a number of addresses that contain “Some County Road” there were cases when the first result returned by Solr returned a false positive. To cater for this, a number of the top results returned from Solr were compared using string similarity metrics. String comparison algorithms/metrics are used to determine the distance or number of changes between two strings (Wagner and Fischer, 1974). The two string comparison methods that were used as part of this experiment were the Levenshtein Distance (Levenshtein, 1966) and Jaro–Winkler Distance (Winkler, 1990). The Levenshtein is computed by calculating the smallest number of single character changes between two strings. The score can be normalised so it produces a value between 0 and 1 by 
\[1 - \frac{\text{number of edits}}{\text{length of the larger string}}.\]

It is very useful for compensating for typos in string matching. The Jaro-Winkler algorithm is used to measure the number of characters in common but also works on the basis that differences at the start of the string are more important than those at end. In research completed by Christen (2006) the Jaro-Winkler method techniques performed quite well across all experiments and was included in this experiment as a result.
Fig. 3.8 illustrates the application that was built as part of this research to map addresses in AIB to a master address database, combining GeoDirectory, Solr and Python.

From test cases it was observed that the application appears to work quite well, especially at matching addresses to Electoral Divisions and Local Authorities which is vital for this research. There are some inaccuracies due to data quality issues with the originating address but that was only when mapping to one specific address in the GeoDirectory database, not electoral division or local authority.

Unfortunately, this research lacked the scope to carry out further analysis on the accuracy of the results. However, the business was so impressed with the results they observed that it is going to carry out an enterprise product investigation and evaluation. Leading on from this research, the Bank they will be taking a sample of 20,000 addresses in AIB and allowing vendors detailed earlier to return their results. These results will then be collated by the business. Testing will be done using crowd sourcing to build confidence intervals to evaluate which product and service offered the best result. An informed data driven decision can then be made if a proprietary address matching
application is needed or the application from this research is accurate and meets the requirements of AIB.

### 3.6 Data for Experiment

As discussed earlier, in the research literature there has been much evidence to support the idea that macro-economic trends such as unemployment and arrears rates at a regional areas are useful features for predicting credit risk and future unperforming SME customers.

The main experiments to be carried out as part of this research will aim to investigate if macro-economic features can improve the prediction model in AIB compared to the results of prediction model based on application and customer behavioural features from the historic scorecard.

5 main categories of experimental features will be created and tested as part of this experiment. These will be delineated in the next subsections of this chapter.

- Personal customer card spending behaviour
- Feature grouping based on Home Loans, Personal Loans, SME Loans
- Central Statistics Office (CSO) features
- SME default behaviour
- Personal Loans and Homeloans default behaviour

#### Personal customer card spending behaviour

One experiment in this research will test to see if customer’s transactional spending behaviour metrics could be useful for predicting if SME customers will default. The intuition for this analysis is that if customers in one area are suffering hardship, their spending habits might be reflective of that and as consumers, this will have an adverse affect on SMEs businesses. To conduct this experiment, transactions will be gathered from the personal customers’ VISA debit card transactional database in AIB. 192 million
transactions from 1.3 million customers will be collected from a 12 month period prior to the observation point (June 2014) in the experiment.

Merchants are assigned what is known as a merchant category code (MCC), which is generally used to classify the primary business of the merchant. These codes have been leveraged in AIB to create a Money Manager Application which rolls these MCC in MCC categories. This application allows personal customers to keep track of their spending behaviour through a combination of reports, visualisations and search functionality. These categories are broken down into parent and child categories for this application. Table 3.2 below outlines these categories.

<table>
<thead>
<tr>
<th>MCC Category</th>
<th>Parent</th>
<th>Child</th>
<th>Spend / Live</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.2</td>
<td>Bills &amp; Utilities</td>
<td>Cable/Satellite TV &amp; Internet</td>
<td>Spend</td>
</tr>
<tr>
<td>2.7</td>
<td>Bills &amp; Utilities</td>
<td>Gas/Electricity/Energy</td>
<td>Live</td>
</tr>
<tr>
<td>3.1</td>
<td>Leisure &amp; Entertainment</td>
<td>Cinema &amp; Theatre</td>
<td>Spend</td>
</tr>
<tr>
<td>4.1</td>
<td>Shopping</td>
<td>Groceries</td>
<td>Live</td>
</tr>
<tr>
<td>4.4</td>
<td>Shopping</td>
<td>Clothing &amp; Accessories</td>
<td>Spend</td>
</tr>
<tr>
<td>5.2</td>
<td>Health &amp; Personal Care</td>
<td>Doctor</td>
<td>Live</td>
</tr>
<tr>
<td>5.6</td>
<td>Health &amp; Personal Care</td>
<td>Hair &amp; Beauty</td>
<td>Spend</td>
</tr>
<tr>
<td>6.2</td>
<td>Household/Home</td>
<td>Household Maintenance</td>
<td>Live</td>
</tr>
<tr>
<td>6.5</td>
<td>Household/Home</td>
<td>Computers &amp; Technology</td>
<td>Spend</td>
</tr>
</tbody>
</table>

Table 3.2: MCC Categories and Spend/Live Categorisation Sample

For this research and experiment, these transactions MCC categories have been assigned *Spend* or *Live* category. Transactions categorised as *Spend* will translate to what has been derived as discretionary spend of AIB customers. This will include transactions like social activities such as going to the cinema, going to bars and clubs and eating out in restaurants (see Table 3.2). Transactions categorised as *Live* will translate to what as been derived as transactions required/associated/needed for customers to live. This will include transactions such as paying bills, shopping/groceries and healthcare.

Transactions for each customer address will be aggregated to electoral division and local authority for each month for the 12 months prior to the observation point. Metrics will be created calculating the difference between time periods to evaluate what trends or patterns are useful for identifying default.

---

Chapter 3. Data

The rates of change between transactional spend will be calculated by using the percentage change equation seen below

\[
\text{Percentage Change} = \frac{X^2 - X^1}{X^1} \times 100
\] (3.1)

Where:
- \(X^1\) = the original variable
- \(X^2\) = the new variable

These features will not be normalised so feature rescaling will be applied to ensure the features’ range value is between 0 and 1. This is achieved through applying the rescaling equation found below

\[
\text{Feature Scaling} = X' = \frac{X - X_{\text{min}}}{X_{\text{max}} - X_{\text{min}}}
\] (3.2)

There will be 68 features created as part of this experiment. The names, data types and description for each feature can be found in Appendix A.

**Personal Loans and Homeloans default behaviour**

As part of the study, an experiment will be carried out to investigate if AIB’s loan and homeloan default ratios and trends could be useful in predicting if SME customers will default. The intuition for this insight is that if there is a high proportion of personal loan and homeloan defaulting in an area, then there could be a relationship with SMEs also defaulting. It was also discussed in literature by Di Pietro and Lusignani (n.d.) how metrics like this were useful for predicting default in Italy.

At the observation point (June 2014), personal loan and homeloan products that were active on the AIB book were analysed. Table 3.3 details the products that were extracted as part of this experiment.
<table>
<thead>
<tr>
<th>Product Category</th>
<th>Product Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>Fixed Loans</td>
<td></td>
</tr>
<tr>
<td>Matrix Loans</td>
<td></td>
</tr>
<tr>
<td>Other Loans</td>
<td></td>
</tr>
<tr>
<td>Branch Advances</td>
<td>Premium Business Rate</td>
</tr>
<tr>
<td></td>
<td>Prime Loans</td>
</tr>
<tr>
<td></td>
<td>Staff Credit Flex</td>
</tr>
<tr>
<td></td>
<td>Suspense Interest</td>
</tr>
<tr>
<td>Buy to Let</td>
<td></td>
</tr>
<tr>
<td>Commercial Mortgages</td>
<td></td>
</tr>
<tr>
<td>Home Loan</td>
<td></td>
</tr>
<tr>
<td>Home Loan</td>
<td>Staff Homeflex</td>
</tr>
<tr>
<td></td>
<td>Standard Mortgages</td>
</tr>
<tr>
<td></td>
<td>Surplus Builder</td>
</tr>
<tr>
<td></td>
<td>Tracker Mortgages</td>
</tr>
</tbody>
</table>

Table 3.3: AIB Personal Loan and Homeloan products

Each product in the Table 3.3 above will be rolled up to its product category. Each product will be associated with a customer and a customer to an address so they can be aggregated to electoral division and local authority. The products are then classified as default or not default so a ratio between the two can be built for each electoral division and local authority. Fig. 3.9 below illustrates the default rates for each local authority for personal loan and homeloan products combined.
Chapter 3. Data

Figure 3.9: Default Ratios for Homeloan and Personal Loans by Local Authority

There will be 6 features created as part of this experiment. The names, data types and description for each features can be observed below in Table 3.4.

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED_HOME_RATIO</td>
<td>Interval</td>
<td>Homeloan Default Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_LOAN_RATIO</td>
<td>Interval</td>
<td>Personal Loan Default Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_TOTAL_RATIO</td>
<td>Interval</td>
<td>Homeloan &amp; Personal Loan Default Ratio for each Electoral Division</td>
</tr>
<tr>
<td>LA_HOME_RATIO</td>
<td>Interval</td>
<td>Homeloan Default Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_LOAN_RATIO</td>
<td>Interval</td>
<td>Personal Loan Default Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_TOTAL_RATIO</td>
<td>Interval</td>
<td>Homeloan &amp; Personal Loan Default Ratio for each Local Authority</td>
</tr>
</tbody>
</table>

Table 3.4: Personal Customers Arrears Ratio features derived for experiment

Central Statistics Office (CSO) Features

Another experiment carried out as part of this research will investigate if data collected as part of the Irish census is indicative of SME default. The census is carried out every 5 years in the Republic of Ireland to paint a picture of the living and social conditions of the population. It provides details to the smallest geographic areas which can be used
for decision making and planning. For example to find what infrastructure and services need to be invested in each areas, for example schools, job training centre and health care services.\(^{14}\)

The data from the census can be organised into 11 themes which are illustrated below in Table 3.5

<table>
<thead>
<tr>
<th>Theme Number</th>
<th>Theme Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Theme 1</td>
<td>Sex, age and marital status</td>
</tr>
<tr>
<td>Theme 2</td>
<td>Migration, ethnicity and religion</td>
</tr>
<tr>
<td>Theme 3</td>
<td>Irish Language</td>
</tr>
<tr>
<td>Theme 4</td>
<td>Families</td>
</tr>
<tr>
<td>Theme 5</td>
<td>Private Household</td>
</tr>
<tr>
<td>Theme 6</td>
<td>Housing</td>
</tr>
<tr>
<td>Theme 7</td>
<td>Communal establishments</td>
</tr>
<tr>
<td>Theme 8</td>
<td>Principal status</td>
</tr>
<tr>
<td>Theme 9</td>
<td>Social class and socio-economic group</td>
</tr>
<tr>
<td>Theme 10</td>
<td>Education</td>
</tr>
<tr>
<td>Theme 11</td>
<td>Commuting</td>
</tr>
<tr>
<td>Theme 12</td>
<td>Disability, careers and general health</td>
</tr>
<tr>
<td>Theme 13</td>
<td>Occupation</td>
</tr>
<tr>
<td>Theme 14</td>
<td>Industries</td>
</tr>
<tr>
<td>Theme 15</td>
<td>PC and internet Access</td>
</tr>
</tbody>
</table>

Table 3.5: Themes of data available in the Irish census 2011

Anecdotally, low unemployment rates and higher education rates would be associated with an area of good economic growth and prosperity. Since the financial crisis of 2008-2009 there has been a huge reduction in the number of construction projects which is said to be heavily linked to higher unemployment rates. As a result, in this experiment features have been created using data provided from the census at local authority and electoral division areas. The features that have been created are lower than secondary level of eduction ratio, manual based occupation ratio, and unemployment ratio by electoral division and local authority.

There will be 6 features created as part of this experiment. The names, data types and description for each features are outlined below in Table 3.6.

\(^{14}\)http://census.ie/
Chapter 3. Data

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED_UNEMPLOYMENT_RATIO</td>
<td>Interval</td>
<td>Employment Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_NON_MANUAL_OCCUPATION_RATIO</td>
<td>Interval</td>
<td>Non Manual Occupation Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_LOWER_THAN_UPPER_SECONDARY_RATIO</td>
<td>Interval</td>
<td>Lower than Secondary Level/Leaving Certificate Ratio for each Electoral Division</td>
</tr>
<tr>
<td>LA_UNEMPLOYMENT_RATIO</td>
<td>Interval</td>
<td>Employment Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_NON_MANUAL_OCCUPATION_RATIO</td>
<td>Interval</td>
<td>Non Manual Occupation Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_LOWER_THAN_UPPER_SECONDARY_RATIO</td>
<td>Interval</td>
<td>Lower than Secondary Level/Leaving Certificate Ratio for each Local Authority</td>
</tr>
</tbody>
</table>

Table 3.6: (CSO) Features derived for experiment

SME Default behaviour

An experiment will be carried out as part of this research to investigate if SME default ratios and behavioural changes could be useful in predicting future SME defaults. One could think of this as a domino effect, where businesses near each other close in sequence due to economic hardship and low consumer confidence in that area. For example, one feature could be to establish if default rates increased or decreased for SME customers at an electoral division or local authority area. It was also discussed in literature by Di Pietro and Lusignani (n.d.) how this was useful for predicting default in Italy.

The data for this experiment was extracted from the SME database in AIB from June 2012 to the June 2014 (Observation Point). Each SME customer’s local authority and electoral division was identified and customers were classified as in default or not in default for each month in the two year period. Features were then built with a focus on identifying local authorities and electoral divisions which had seen higher default rates in the two years prior to the observation point.

The majority of the features will be built using the percentage change formula seen above in Equation 3.1 and the rescaling formula Equation 3.2. It will also include the ratio between default and non-default and count of default for each local authority and electoral division.

There will be 20 features created as part of this experiment. The names, data types and description for each features can be found below in Table 3.7.
Table 3.7: SME default behaviour features derived for experiment

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA_PERCENT_2006_2014</td>
<td>Interval</td>
<td>SME Default Ratio for each local authority at June 2014</td>
</tr>
<tr>
<td>ED_PERCENT_2006_2014</td>
<td>Interval</td>
<td>SME Default Ratio for each electoral division at June 2014</td>
</tr>
<tr>
<td>LA_CNT_2006_2014</td>
<td>Interval</td>
<td># SME Defaults for each local authority at June 2014</td>
</tr>
<tr>
<td>ED_CNT_2006_2014</td>
<td>Interval</td>
<td># SME Defaults for each electoral division at June 2014</td>
</tr>
<tr>
<td>DIFF_12_2013_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2013 for each local authority</td>
</tr>
<tr>
<td>DIFF_06_2013_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2013 for each local authority</td>
</tr>
<tr>
<td>DIFF_12_2012_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2012 for each local authority</td>
</tr>
<tr>
<td>DIFF_06_2012_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2012 for each local authority</td>
</tr>
<tr>
<td>DIFF_12_2013_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2013 for each electoral division</td>
</tr>
<tr>
<td>DIFF_06_2013_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2013 for each electoral division</td>
</tr>
<tr>
<td>DIFF_12_2012_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2012 for each electoral division</td>
</tr>
<tr>
<td>DIFF_06_2012_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2012 for each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2013_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2013 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2013_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2013 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2012_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2012 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2012_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2012 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2013_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2013 by each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2013_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2013 by each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2012_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2012 by each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2012_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2012 by each electoral division</td>
</tr>
</tbody>
</table>

Feature grouping based on Home Loans, Personal Loans, SME Loans

As reviewed in the literature in Section 2.10, binning is a useful method of transforming interval based features into categorical features. This has a number of benefits, including simplifying the structure of the data into nominal and ordinal based features and interval based on estimating the group’s weight of evidence (WoE). Binning increases the robustness of models by generalising models for unseen data, thus reducing the chances of over-fitting. As discussed in the literature it also has the capability to incorporate missing values and other extreme outliers which cause instability in the model.

The features to be binned as part of this experiment are based on features already seen in this section. Homeloan, personal loan and SME default ratios at the observation point (June 2014) will be binned in attempt to build a robust and simplified feature set. This experiment will test the nominal/ordinal outputs and WoE outputs in the predictive modelling stage to see if either improve the prediction of the baseline model.

There will be 16 features created as part of this experiment. The names, data types and description for each features can be found below in Table 3.8.
### Table 3.8: Binned/Grouped features based on default ratios of SME, Home and Personal Loans by Electoral Division and Local Authority

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUPED_ED_ID_SME_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for electoral division default ratios for SME loans</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_SME_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on electoral division default ratios for SME loans</td>
</tr>
<tr>
<td>GROUPED_LA_ID_SME_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authority default ratios for SME loans</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_SME_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on local authority default ratios for SME loans</td>
</tr>
<tr>
<td>GROUPED_LA_ID_BOTH_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authority default ratios for Homeloan &amp; Personal loans</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_BOTH_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on local authority default ratios for Homeloans and Personal Loans</td>
</tr>
<tr>
<td>GROUPED_ED_ID_BOTH_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for electoral division default ratios for Homeloans &amp; Personal loans</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_BOTH_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on electoral division default ratios for Homeloans and Personal Loans</td>
</tr>
<tr>
<td>GROUPED_ED_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for electoral division default ratios for Personal loan only</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on electoral division default ratios for Personal Loans</td>
</tr>
<tr>
<td>GROUPED_LA_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authority default ratios for Personal loan only</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on local authority default ratios for Personal Loans</td>
</tr>
<tr>
<td>GROUPED_ED_ID_HOMELOANS_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for electoral division default ratios for Homeloans only</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_HOMELOANS_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on electoral division default ratios for Homeloans</td>
</tr>
<tr>
<td>GROUPED_LA_ID_HOMELOANS_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authority default ratios for Homeloans only</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_HOMELOANS_ARREARS</td>
<td>Interval</td>
<td>Weight for each bin value based on local authority default ratios for Homeloans</td>
</tr>
</tbody>
</table>

#### 3.7 Generating the ABTs

The data that will be collected and processed as part of this research were outlined in the previous sections. It can be seen in Fig. 3.10 how each internal data source in AIB will have to be mapped to an address through the address matching application.
When the process is finished there will be two ABTs generated for SME customers based on features from historic scorecards and two ABTs that also incorporate experimental features based on the electoral division and local authority.

<table>
<thead>
<tr>
<th>Model</th>
<th># Numeric</th>
<th># Nominal</th>
<th># Observations</th>
<th># Good</th>
<th># Bad</th>
<th>Good:Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Delinquency</td>
<td>119(Previously 11)</td>
<td>8(Previously 0)</td>
<td>2,926</td>
<td>2,198</td>
<td>738</td>
<td>75:25</td>
</tr>
<tr>
<td>No Previous Delinquency</td>
<td>117(Previously 9)</td>
<td>8(Previously 0)</td>
<td>24,156</td>
<td>23,505</td>
<td>651</td>
<td>97:03</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td></td>
<td>27,082</td>
<td>25,703</td>
<td>1,389</td>
<td>95:05</td>
</tr>
</tbody>
</table>

Table 3.9: Characteristics of datasets to be used in the experiment to evaluate macro-economic features

# Numeric refers to the number of continuous features
# Nominal refers to the number of categorical features

3.8 Software Used

A number of software tools and applications were required in this research. These have been broken down into four main subsections; address matching, data wrangling to
create ABT, visualisations and modelling. All the software used in this research was open source, apart from SAS which is used for the benchmark evaluation experiment in Section 4.2.

**Address Matching**

A combination of Apache Solr\(^{15}\) and Python\(^{16}\) was used to match AIB’s addresses to an ED/LA in GDD. Solr is an open source web application enterprise search engine, written in Java, which is a wrapper around Apache Lucene\(^{17}\). Combined they provide a reliable, fast, scalable platform capable of providing distributed indexing which can then be used for searching or navigation. Solr was used to index GeoDirectory Database which then allows it to be searched. Python is a high-level, general purpose programming language which can be used to build both large and small scale programs. Python, like Solr, is open source and freely available. One of its most attractive and best characteristics is that it is easy to read and use.

**Data Wrangling to create ABT**

Data scientists and analysts spend the majority of their time data wrangling. Data wrangling is a time consuming, mundane process used to collect and prepare data prior to being explored for useful information. In the experiment for this study, a variety of data types and data sources were used. Data from the AIB EDW, GDD in Solr served in JSON, CSO data in flat file are just some of the sources and types. R\(^{18}\), an open source programming language has much more emphasis on statistical computing. It also has many libraries available for processing and data manipulation which are available in the CRAN\(^{19}\) repository. The most useful package used during this process was reshape\(^{20}\). Reshape allows one to easily restructure, transpose and aggregate data.

---

\(^{15}\)http://lucene.apache.org/solr/

\(^{16}\)https://www.python.org/

\(^{17}\)https://lucene.apache.org/core/

\(^{18}\)https://www.r-project.org/

\(^{19}\)https://cran.r-project.org/

\(^{20}\)https://cran.r-project.org/web/packages/reshape/index.html
Chapter 3. Data

Visualisations

R is also a very strong programming language in terms of creating beautiful visualisations, so it will be used throughout this paper. One package used through the study was `ggplot`\(^{21}\).

For Geographic Information Systems (GIS) that have been created as part of this research QGIS\(^{22}\) has been used. QGIS formally known as Quantum GIS is a cross platform open source GIS application allowing one to create maps compiled with data for information sharing and analysis. It is the challenger to the proprietary application ArcGIS\(^{23}\) as both provide a very similar interface and functionality for creating and sharing maps.

Modelling

SQL was used to identify customers to be used for predictions and generate the target class. The models and experiments performed in this research were built in R and SAS. SAS is a proprietary software and the tool of choice by the modelling teams in AIB. Anecdotally, SAS is a legacy system in financial institutions, it is what people are used to using but also there is a for-profit corporation vetting the code for its customers, offering customer service and support which corporations are used to. SAS offers a graphical interfaces which means users do not have to enter code, but this can be complemented using the SAS programming language. SAS is excellent for building predictive models resulting in good time to value. SAS Enterprise Miner includes the following components Time Series, Variable Clustering, Cluster, Interactive Binning, Principal Components, AutoNeural, DMNeural, DMine Regression, Gradient Boosting, Ensemble, and Text Mining.

R is a very strong competitor to SAS in this space. Because of its open source nature there are many libraries available for building predictive models. For example one popular package `caret`\(^{24}\) contains many models and it continues to grow.

The experiment in Section 4.2 will be built using SAS. Experiments in later sections will be completed using R. This was done so as to leverage the industry supported tool

\(^{21}\)https://cran.r-project.org/web/packages/ggplot2/index.html
\(^{22}\)http://www.qgis.org/en/site/
\(^{23}\)https://www.arcgis.com/features/
\(^{24}\)https://cran.r-project.org/web/packages/caret/index.html
for decisions such as model selection and performance measure selection. For further experiments it was found that R had a much quicker time to value for manipulation of data, feature selection and model comparison. For R the data was being stored locally on the author’s machine whereas SAS runs on an industry server in AIB which is heavily administrated.

### 3.9 Conclusions

This chapter has discussed what data will be used as part of this research project.

Divided into two main sections, the chapter first details on which small-medium enterprises (SME/SMEs) customers the prediction model will be trained and the time-period on which the prediction will be made. The building of a benchmark model using features selected from a historic AIB credit scorecard which will be used for comparisons carried out as part of the research experiment was delineated.

The second main section discussed the macro-economic features that were generated as part of this research project. These features were generated from internal sources in AIB and open data from the census. An address-matching application was designed and built as part of the macro-economic feature generation stage which mapped AIB customer address data to geographic regions in Ireland. AIB were so impressed by the results that a project to evaluate the application will be conducted to see if it can be leveraged for matching addresses in AIB to the Eircode system.

A methodology for generating multiple analytical base tables is detailed which covers where the data is sourced and where data wrangling is carried out.

Finally, software used in the data generation stage is outlined.
Chapter 4

Design and Methodology

4.1 Introduction

This chapter will present the design and methodology of the experiments to be undertaken as part of this research. The aim of this research is to establish if including macro-economic features while training a SME credit scoring model will return better results than a model trained using the benchmark features from a historic scorecard.

This chapter comprises of two main sections. The first will consist of an exploratory evaluation, building a predictive model and evaluating the benchmark features for previous delinquency and no previous delinquency datasets. Models and performance measures will be assessed as part of this experiment, the dataset will be split into training, validation and test sets. The approach will involve training multiple models based on features selected for a historic credit scorecard in order to evaluate which is the best. Each model will be assessed to ascertain the strength of each model, enabling model selection for future experiments in this research. This section will also explore and evaluate the results from performance measures applied to establish the accuracy of the model. These performance measures will be used to elicit the overall accuracy of the model, how well it generalises and also accuracy over a specific threshold.

The second part of the experiment will involve assessing the use of macro-economic features when building a predictive model. Another baseline model will be built using the model selected at the model evaluation stage. Modelling after feature selection and coarse selection modelling will be outlined. Following on from these experiments, this
section delineates an approach to tackling the class imbalance issue by using oversampling of the minority class.

4.2 Benchmark Features Evaluation

When building a predictive model, it is useful to have a baseline or benchmark model from which to make model comparisons. The initial task is to build a model where the results/predictions are better than the no information rate. The no information rate is taken to be the biggest class percentage in the data to be modelled. The accuracy of the model must be better than this rate.

However, for the purposes of this study, it would be redundant to build a model that is better than the no information rate as there are already industry models in existence in AIB for predicting arrears which can be leveraged. For this research the, the Risk team in AIB has provided two feature sets which have been used historically for predicting arrears. The team have segmented the data into two groups, Previous Delinquency and No Previous Delinquency. As discussed in Section 3.3, this was done because features that detailed whether a customer had been in default previously would be dominant in a training model on the full population.

Therefore, two models will be built as benchmarks which will be compared to the results from later experiments. One model will be based on a feature set for customers who have been delinquent in the past and the other feature set for customers who have not been delinquent in the past. The customers will be modelled with these features first and results will be recorded. In later sections, as part of the experiment, location based features will be added to be modelling process, with the aim of establishing that these features will be statistically significant for predicting SME arrears.

In Section 2.12, many model validation methods were discussed, including LOOCV, k-fold validation and hold-out validation method. Due to practicability and reproducibility, it was decided that hold-out method with training/validation/test dataset would be the best to assess model results using the performance measures. As mentioned previously, at the observation point, June 2014, SME customers who were not in default were selected. There were 27,082 customers in these experiments which is divided into a previous delinquency and no previous delinquency dataset. Stratified sampling was
used on the target class to build the training, validation and test datasets. The split will be broken down as 60% training, 20% validation and 20% testing. The validation will be used for setting the threshold parameter for assessing the performance of the model. The breakdown for these partitions can be observed in Table 4.1 below.

Note: The data partitions will be consistent across all tests to mitigate the risk of misleading results.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th># Bad</th>
<th># Good</th>
<th># Observations</th>
<th>Good:Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Delinquency</td>
<td>Training</td>
<td>436</td>
<td>1,318</td>
<td>1,754</td>
<td>75:25</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>146</td>
<td>440</td>
<td>586</td>
<td>75:25</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>146</td>
<td>440</td>
<td>586</td>
<td>75:25</td>
</tr>
<tr>
<td>Previous Delinquency</td>
<td>Total</td>
<td>728</td>
<td>2,198</td>
<td>2,926</td>
<td>75:25</td>
</tr>
<tr>
<td>No Previous Delinquency</td>
<td>Training</td>
<td>391</td>
<td>14,103</td>
<td>14,494</td>
<td>97:03</td>
</tr>
<tr>
<td></td>
<td>Validation</td>
<td>130</td>
<td>4,701</td>
<td>4,831</td>
<td>97:03</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>130</td>
<td>4,701</td>
<td>4,831</td>
<td>97:03</td>
</tr>
<tr>
<td>No Previous Delinquency</td>
<td>Total</td>
<td>651</td>
<td>23,703</td>
<td>24,156</td>
<td>97:03</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1,379</td>
<td>25,703</td>
<td>27,082</td>
<td>95:05</td>
</tr>
</tbody>
</table>

Table 4.1: Breakdown Holdout Training/Validation/Test Dataset for Benchmark Evaluation

It can be observed in Table 4.1 that the Good:Bad ratio is consistent for each partition because stratified sampling based on the target feature ensures each partition shares the same homogeneous properties as the full dataset.

Five different predictive models will be trained as part of this experiment. These were discussed previously in Section 2.8. Typically in AIB and industry, logistic regression is the predictive model of choice. This experiment will evaluate if there is any significant difference between the results from a logistic regression compared with a decision tree, neural network, gradient boosting, and SVM models. As discussed in the research, these are some of the most common algorithms used in predictive modelling.

Once each of the models have been trained, they will be assessed for their performance and accuracy. Based on the literature and industry practice, the AUC is the performance measure of choice for evaluating how well the model performed across all test observations. When combined with the ROC chart, it allows you identify if the model generalised well or if the model has over-fitted the training data as well as the overall
accuracy of the model. However, AUC does not facilitate the identification of the threshold to be chosen to categorise predictions into ‘goods’ and ‘bads’. Therefore, statistics such as the minimum misclassification rate, K-S statistic, equal precision equal recall (EPER), lift and the default threshold will be evaluated to identify the optimal threshold to split the test results.

The evaluation process used for model and performance measure is illustrated below in Fig. 4.1.

Figure 4.1: Overview of Benchmark Evaluation Process for Model and Threshold Selection

Previous Delinquency Dataset Benchmark Evaluation

The dataset to model customers who have been in default in the past contains 11 features. For privacy reasons, the names and descriptions of these features could not be disclosed. There is an imbalance in the Previous Delinquency dataset; of the total number of customers to be modelled, approximately 25% of customers are in default at the end of outcome window (See Table 4.1). This default rate is derived from customers in this analysis and is not reflective of the enterprise default rate. The results for the benchmark models can be seen in Table 4.2.
Table 4.2 illustrates the model with the best result. Highlighted in green in Fig. 4.2, it is the Gradient Boosting Model(Ada), which achieves the highest validation and test AUC. We can see above Table 4.2 and Fig. 4.2 below that the Gradient Boosting Model generalises quite well across the training, validation and testing partitions.

Table 4.2: Previous Delinquency Benchmark Results

<table>
<thead>
<tr>
<th>Model</th>
<th>Train AUC</th>
<th>Train GINI</th>
<th>Valid AUC</th>
<th>Valid GINI</th>
<th>Test AUC</th>
<th>Test GINI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Gradient Boosting</td>
<td>0.655</td>
<td>0.331</td>
<td>0.681</td>
<td>0.344</td>
<td>0.597</td>
<td>0.195</td>
</tr>
<tr>
<td>Regression</td>
<td>0.65</td>
<td>0.301</td>
<td>0.672</td>
<td>0.344</td>
<td>0.597</td>
<td>0.195</td>
</tr>
<tr>
<td>AutoNeural Network</td>
<td>0.65</td>
<td>0.301</td>
<td>0.672</td>
<td>0.344</td>
<td>0.597</td>
<td>0.195</td>
</tr>
<tr>
<td>Regression Backstep</td>
<td>0.643</td>
<td>0.287</td>
<td>0.661</td>
<td>0.323</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Regression Forward Step</td>
<td>0.643</td>
<td>0.287</td>
<td>0.661</td>
<td>0.323</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>Regression Both</td>
<td>0.643</td>
<td>0.287</td>
<td>0.661</td>
<td>0.323</td>
<td>0.6</td>
<td>0.2</td>
</tr>
<tr>
<td>SVM Polynomial</td>
<td>0.654</td>
<td>0.308</td>
<td>0.62</td>
<td>0.241</td>
<td>0.593</td>
<td>0.186</td>
</tr>
<tr>
<td>SVM Radial Basis Fn</td>
<td>0.812</td>
<td>0.624</td>
<td>0.6</td>
<td>0.2</td>
<td>0.619</td>
<td>0.238</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.626</td>
<td>0.252</td>
<td>0.588</td>
<td>0.176</td>
<td>0.55</td>
<td>0.1</td>
</tr>
<tr>
<td>SVM Sigmoid</td>
<td>0.492</td>
<td>-0.016</td>
<td>0.511</td>
<td>0.241</td>
<td>0.023</td>
<td>-0.018</td>
</tr>
</tbody>
</table>

It is evident from Table 4.2 that there no one stand-out model and there is an argument that logistic regression should be continued to be the approach of choice. It can be observed that the SVM Radial Basis Fn model does not generalise well. It appears to
have completely over-fitted the training partition with an AUC training = 0.812 which drops to 0.60 and 0.619 for the validation and testing partitions respectively. Overall, most of the models appear to be predictive and have generalised quite well. However, there may be a case for investigating and removing the SVM Radial Basis Fn, Decision Tree and SVM Sigmoid as these appear not to be predictive, have not generalised well or have over-fitted the training data partition.

Although the AUC is very useful for detailing how well the prediction performed over all possible thresholds, it is not very useful splitting the predictions into ‘goods’ and ‘bads’. In an ideal scenario, the target class distribution would be 50:50 and this would allow the baseline/default threshold of 0.50. When there is an imbalance in the dataset, many thresholds need to be evaluated to test which one meets the business objective.

Recall measures the percentage of defaulters (positive class) that were identified correctly. Accuracy measures what percentage of the predictions made were correct. Specificity measures the percentage of the non-defaulter (negative class) that were identified correctly. Precision measures the percentage of the positive predictions made that were correctly identified. Balanced accuracy is the average of the sum of recall and specificity. All of these performance measures are based on a single threshold in the predictions made i.e. a cut-off point.

Selecting the cut-off point is crucial for any business objective. In predicting defaults, the focus is on maximising recall where possible. However, maximising this performance measure is not possible because all the ‘good’ customers will also be predicted ‘bad’. Balanced accuracy is a useful performance measure as it uses how many of the negative class and positive class are being classified correctly. However, it is an average therefore the results could be skewed by one performance measure performing well. Therefore, one must check recall and specificity, to ensure there is no bias. This experiment will aim to maximise balanced accuracy, recall and specificity.

An experiment will be generated for the previous delinquency and no previous delinquency dataset, where the threshold will be selected based on the validation dataset and then evaluated on the test dataset. As a baseline, a default cut-off of 0.50 will be included in the experiment. The minimum misclassification rate, Kolgorov-Smirnov (K-S) statistic and event precision event recall (EPER) methods will be used to evaluate cut-off points. K-S statistic measures at what cut-off the maximum difference between
the cumulative positive and negative distributions of the predicted positive and negative class. EPER measures at what value the cut-off precision and recall measures are the same. The min misclassification rate is the cut-off where the model makes the most correct prediction.

Results for the previous delinquency data can be observed below in Table 4.3.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Method</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>Default (Valid)</td>
<td>0.763</td>
<td>0.055</td>
<td>0.998</td>
<td>0.526</td>
</tr>
<tr>
<td>0.5</td>
<td>Default (Test)</td>
<td>0.765</td>
<td>0.062</td>
<td>0.998</td>
<td>0.53</td>
</tr>
<tr>
<td>0.48</td>
<td>Min MR (Valid)</td>
<td>0.758</td>
<td>0.062</td>
<td>0.989</td>
<td>0.526</td>
</tr>
<tr>
<td>0.48</td>
<td>Min MR (Test)</td>
<td>0.765</td>
<td>0.068</td>
<td>0.995</td>
<td>0.532</td>
</tr>
<tr>
<td>0.25</td>
<td>K-S* (Valid)</td>
<td>0.666</td>
<td>0.548</td>
<td>0.705</td>
<td>0.627</td>
</tr>
<tr>
<td>0.25</td>
<td>K-S* (Test)</td>
<td>0.630</td>
<td>0.507</td>
<td>0.670</td>
<td>0.586</td>
</tr>
<tr>
<td>0.29</td>
<td>EPER* (Valid)</td>
<td>0.705</td>
<td>0.370</td>
<td>0.816</td>
<td>0.593</td>
</tr>
<tr>
<td>0.29</td>
<td>EPER* (Test)</td>
<td>0.688</td>
<td>0.342</td>
<td>0.802</td>
<td>0.572</td>
</tr>
</tbody>
</table>

Table 4.3: Previous Delinquency Cut-off Results

- K-S = Optimal threshold for Kolgororov-Smirnov Statistic
- Min MR = Optimal threshold for Min Mis-classification Rate
- EPER = Optimal threshold for Event Precision Equals Recall
- BA = Balanced Accuracy = \( \frac{\text{Recall} + \text{Specificity}}{2} \)

The validation data is used to choose the cut-off for the measures and then the test data is then evaluate performance on unseen data. Optimal values are coloured in yellow in Table 4.3.

The min mis-classification rate calculates the accuracy of model, but because the model is so heavily imbalanced, the majority negative class (good customers) is skewing the results. It can be seen that the cut-off that gives the min misclassification rate scores accuracy very well, but performs terribly on the recall. As a result, this will be discounted.

The EPER cut-off derives the threshold where the precision rate and recall rate intersect. This method performs extremely well on the specificity, while also performing at an acceptable level on the recall.

The KS statistic gives the threshold maximising the separation between ‘good’ and ‘bads’. It scores recall and balanced accuracy highly, while specificity also performs
well. For this reason, the threshold to be used when building the benchmark model for previous delinquency will be where the cut-off value is equal to 0.25.

**No Previous Delinquency**

In the dataset to model customers who have not been in default in the past, there are 9 features. For privacy reasons, the names and descriptions of these features cannot not be disclosed. There is an imbalance in the No Previous Delinquency dataset; of the total customers to be be modelled, approximately 2.7% of customers are in default at the end of outcome window (see Table 4.1). This default rate is derived from customers in this analysis and is not reflective of the enterprise default rate. The results for the baseline benchmarks models are as outlined in Table 4.4.

<table>
<thead>
<tr>
<th>Model</th>
<th>Train AUC</th>
<th>Train GINI</th>
<th>Valid AUC</th>
<th>Valid GINI</th>
<th>Test AUC</th>
<th>Test GINI</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regression</td>
<td>0.695</td>
<td>0.389</td>
<td>0.71</td>
<td>0.419</td>
<td>0.677</td>
<td>0.354</td>
</tr>
<tr>
<td>Regression Backstep</td>
<td>0.691</td>
<td>0.381</td>
<td>0.706</td>
<td>0.413</td>
<td>0.678</td>
<td>0.357</td>
</tr>
<tr>
<td>Regression Forward Step</td>
<td>0.691</td>
<td>0.381</td>
<td>0.706</td>
<td>0.413</td>
<td>0.678</td>
<td>0.357</td>
</tr>
<tr>
<td>Regression Both</td>
<td>0.691</td>
<td>0.381</td>
<td>0.706</td>
<td>0.413</td>
<td>0.678</td>
<td>0.357</td>
</tr>
<tr>
<td>Gradient Boosting</td>
<td>0.653</td>
<td>0.305</td>
<td>0.683</td>
<td>0.366</td>
<td>0.688</td>
<td>0.336</td>
</tr>
<tr>
<td>SVM Radial Fn</td>
<td>0.591</td>
<td>0.182</td>
<td>0.558</td>
<td>0.116</td>
<td>0.512</td>
<td>0.025</td>
</tr>
<tr>
<td>SVM Sigmoid</td>
<td>0.605</td>
<td>0.21</td>
<td>0.549</td>
<td>0.099</td>
<td>0.588</td>
<td>0.176</td>
</tr>
<tr>
<td>AutoNeural Network</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>Decision Tree</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
<td>0.5</td>
<td>0</td>
</tr>
<tr>
<td>SVM Polynomial</td>
<td>0.477</td>
<td>-0.046</td>
<td>0.497</td>
<td>-0.007</td>
<td>0.487</td>
<td>-0.026</td>
</tr>
</tbody>
</table>

*Table 4.4: No Previous Delinquency Bench Mark Results*

Highlighted in green in Table 4.4, Logistic Regression is the model with the best result, with the highest validation and test AUC. From Table 4.4 above and Fig. 4.3 below, it is evident that the Logistic Regression model generalises quite well across the training, validation and testing partitions.
It can be observed that the majority of the models generalise quite well and are predictive. However, it is also evident from Table 4.4 and Fig. 4.3, that the AutoNeural Network, Decision Tree and SVM Polynomial models are not at all predictive. Therefore, these models will be removed from any further analysis. Of the remaining models, logistic regression performed the best; just as it did in the previous delinquency dataset element of this experiment.

The next step is to evaluate a threshold for the no previous delinquency dataset. Results for the no previous delinquency data can observed below in Table. 4.5.
The validation data is used to choose the cut-off for the measures and then the test data is used to evaluate performance on unseen data. Optimal values are coloured in yellow in Table 4.5.

As in the previous experiment, the threshold for the K-S statistic performs significantly better when analysing the recall, specificity and balanced accuracy. For this reason, the threshold to be used when building the benchmark model for the no previous delinquency dataset will be where the cut-off value is equal to 0.03.

### 4.3 Methodology for Evaluating Macro-Economic Features

Leading on from the results in Section 4.2, this part of the chapter will delineate the design of the experiment to evaluate macro-economic features created as part of this research.

Fig. 4.4 below illustrates at a high level, the design of this experiment.

<table>
<thead>
<tr>
<th>Cut-off</th>
<th>Method</th>
<th>Accuracy</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.5</td>
<td>Default (Valid)</td>
<td>0.972</td>
<td>0.000</td>
<td>1.000</td>
<td>0.5</td>
</tr>
<tr>
<td>0.5</td>
<td>Default (Test)</td>
<td>0.972</td>
<td>0.000</td>
<td>1.000</td>
<td>0.5</td>
</tr>
<tr>
<td>0.21</td>
<td>Min MR* (Valid)</td>
<td>0.972</td>
<td>0.000</td>
<td>0.999</td>
<td>0.495</td>
</tr>
<tr>
<td>0.21</td>
<td>Min MR* (Test)</td>
<td>0.972</td>
<td>0.008</td>
<td>0.99</td>
<td>0.499</td>
</tr>
<tr>
<td>0.03</td>
<td>K-S (Valid)</td>
<td>0.743</td>
<td>0.538</td>
<td>0.749</td>
<td>0.644</td>
</tr>
<tr>
<td>0.03</td>
<td>K-S (Test)</td>
<td>0.739</td>
<td>0.515</td>
<td>0.745</td>
<td>0.63</td>
</tr>
<tr>
<td>0.08</td>
<td>EPER* (Valid)</td>
<td>0.942</td>
<td>0.146</td>
<td>0.964</td>
<td>0.556</td>
</tr>
<tr>
<td>0.08</td>
<td>EPER* (Test)</td>
<td>0.942</td>
<td>0.162</td>
<td>0.963</td>
<td>0.563</td>
</tr>
</tbody>
</table>

Table 4.5: No Previous Delinquency Cut-off Results

K-S = Optimal threshold for Kologorov-Smirnov Statistic

MR = Optimal threshold for Mis-classification Rate

EPER = Optimal threshold for Event Precision Equals Recall

BA = Balanced Accuracy = \( \frac{\text{Recall} + \text{Specificity}}{2} \)
As discussed in Chapter 3, data for this experiment will include features used in a historical credit scorecard model in AIB (see Section 3.3), along with macro-economic features built from transaction spend, census metrics, default ratios and trends (see Section 3.6).

Stratified sampling will be applied to the target class to build the training and test datasets. The split will be 70% training and 30% test, as recommended in research conducted by Siddiqi (2012). Since parameters for the model have been tuned using the benchmark evaluation, there is no longer a requirement for a validation dataset. A baseline benchmark model will be built using this 70/30 split of data which will be evaluated using the AUC and balanced accuracy using the threshold from the KS statistic discussed in the previous section. Due to results outlined in earlier sections, logistic regression is the algorithm of choice.

Once the benchmark model has been built, the next phase will be to evaluate the macro-economic features. As part of the study, 116 features were created. As part of the study,
116 features were created. Models can suffer from the curse of dimensionality where too many features are trained in the model (see Section 2.9). This can cause over-fitting and the resulting model may not generalise well when making predictions on unseen data. As noted, research suggests as a guideline, a model should encompass between 8 and 20 features (Thomas, 2009; Mays, 2004).

Therefore, feature selection will be used in this experiment to analyse what features have the strongest relationship with the target feature. Two methods of features selection will be explored; correlation feature based selection and information gain. The features demonstrating the strongest relationship to the target feature will be added for model training and evaluation. The correlation based feature selection will be run over the combined datasets (previous delinquency and no previous delinquency). The information gain feature selection will be run twice, once over each homogeneous dataset separately (previous delinquency and no previous delinquency). Variable importance will be applied, using the random forests method. The random forest feature importance is used to generate an indicator of the importance of each feature in the training dataset. This is done by generating 1000 decision trees with subsets of the training dataset features. Collating these results will identify what features were of greatest importance at the decision node of each tree.

Next, coarse selection will be applied to try and reduce the macro-economic features to a manageable level. As reviewed in Section 2.10, this method will transform all the continuous macro-economic features into nominal/ordinal features. These nominal features will have a small number of values. As discussed earlier, this allows the model to become more robust, mitigating the risk of over-fitting and also allowing it handle missing values for features. As already outlined, the information gain of each feature will then be calculated and the binned features demonstrating the strongest relationship to the target class will be added for model training and evaluation.

Cognisant that the distribution of the number of ‘good’ and ‘bad’ customers is very imbalanced in the dataset for prediction, random over sampling will be used to tackle the issue. This method creates new instances of the minority dataset on which the model can be trained. This method will be applied to the datasets which include macro-economic features in order to evaluate if these models performed better than imbalanced tests.
It must be noted that all feature and coarse selection were carried out on the same training datasets as the benchmark model. Also it is important to ensure that feature and coarse selection must only be carried out on the training dataset; otherwise the model could end up over-fitting the test data which may cause misleading results. All models will be trained, tested and evaluated based on results discussed in Section 4.2. All results will then be documented for comparison with the benchmark model.

### 4.4 Conclusions

This chapter focussed on two main topics. The first of which centred on the design and build of an experiment to evaluate the best algorithm to create the benchmark prediction model. Decision trees, logistic regression, auto neural network, Support Vector Machine (SVM) Polynomial, SVM Radial Basis Function, SVM Sigmoid, and Gradient Boosting were evaluated. AUC was used to establish the accuracy of each model across all possible thresholds. It was found that SVM Sigmoid model performed poorly for both the previous delinquency and no previous delinquency datasets so it was discarded. The best performing algorithms were logistic regression and gradient boosting; in both datasets they emerged as the best and second best. Due to the wide use and support for logistic regression in credit scoring, it was decided that it would be chosen as the algorithm for modelling in the remaining experiments.

AUC is the performance measure of choice for assessing how a model performed over all possible thresholds but it does not detail the threshold to be chosen to best categorise the predictions into ‘goods’ and ‘bads’. An experiment was conducted to evaluate what threshold should be selected to maximise performance measures such as recall, specificity and balanced accuracy. The threshold offered by using the K-S statistic gave the best overall results for recall, specificity and balanced accuracy across the previous delinquency and no previous delinquency dataset.

The second focus of this chapter centred on the experiment design for evaluating the predictive capability of macro-economic features in modelling credit risk for small and medium enterprises. This experiment will incorporate two benchmark models; one for the previous delinquency dataset and one for the no previous delinquency dataset. Macro-economic features will be put through various feature selection processes in an
effort to identify features which could be predictive in the model. There will also be an experiment to address the target feature imbalance in these datasets. Random sampling of the minority class will be used to achieve this.
Chapter 5

Implementation and Evaluation

5.1 Introduction

This chapter will present the implementation of the experiments to evaluate the use of macro-economic features for predicting SME defaults. The predictive capability of customers’ spending behaviour, personal customer default rates, SME default rates and Census data will be analysed and discussed.

The experiments in this research are sequential meaning results from one experiment are used in the following experiment. Therefore results will also be discussed and evaluated in this section.

A benchmark predictive model will be built in SAS and R based on historical scorecard features. This will be used to make fair comparisons of the results from the experiments.

5.2 Data Exploration

This section will explore some of the trends and patterns of SME default rates of customers in the dataset. We will visualise SME defaults trends from June 2012 - June 2015 by looking at the number of SME customers in default each month and the percentage of customers in default each month to identify any noteworthy findings.

Finally we will visualise the percentage of SME customers in default by electoral division and local authority at the observation point to see if there is reason to believe that
location has any obvious correlations with default rates which will be done using a geographic information system (GIS) application.

**Figure 5.1:** SME Customers Default/Performing June 2012 - June 2015

- **x-axis =** Year-Month
- **y-axis =** # SME Customers
- **Red Trend Line =** # SME Customers in Default each month
- **Blue Trend Line =** # SME Customers in Performing each month

Fig. 5.1 shows the number of SME customers that are performing or are defaulting each month from June 2012-June 2015. The data appears to be relatively well behaved. You can see to the right of the chart that the number of SME customers performing is increasing while at the same time the number of customers in default is dropping.

**Figure 5.2:** Percentage of SME Customers in Default Each Month

- **June 2012 - June 2015**
- **x-axis =** Year-Month
- **y-axis =** % of SME Customers in Default
- **Black Trend Line =** % SME Customers in Default Each Month
Fig. 5.2 shows the percentage of SME customers that are in default each month from June 2012 - June 2015. The trend of default is more obvious in this visualisation compared to Fig. 5.1. You can see that the lowest percentage of defaulting customers occurs in June 2012 (left of the chart) where default percentage is just over 17.8%. This percentage increases over time reaching a maximum at July 2014 (Observation Point) with a percentage of 22.5%. By June 2015 the SME default has fallen significantly to just over 18%. This could make predictions using macro-economic features challenging as in the trained dataset the percentage of SME default is rising while a prediction is being made when the percentage of SME default is falling.

Fig. 5.3 below is a GIS application of the % SME Customers in Default by County and Electoral Division at the observation point (June 2014). Regions that have a small percentage of SME customers in default are coloured in white while regions that have a high percentage of default are coloured a dark red.

The results of this visualisation are quite interesting and backup the hypothesis that
location may be an important contributory risk factor in SME default rates in Ireland. Firstly looking at Fig. 5.3a we can see that the local authorities with a high percentage of default cluster to the east coast of Ireland with Cavan, Meath, Louth, Dublin performing the worst.

It is also evident that default rates in cities are greater than the counties in which they’re located. This is true of Galway, Limerick, Cork and Waterford. This phenomenon is particularly evident in Waterford where a huge disparity exists between default rates in the city and county.

Fig. 5.3b is a little bit more difficult to draw conclusions from due to the number of electoral division in Ireland. It does demonstrate that within each local authority there is a large amount of variation between the percentage default of a electoral divisions. Focusing on the Dublin area in particular in Fig. 5.3b it can be seen that large differences exist between electoral divisions with some performing well and others performing poorly.

The results from this section have been very useful. They have proved that the data is well behaved in Fig. 5.1. It has identified that falling SME default percentages from the observation point could provide a challenge Fig. 5.2 in making predictions. Results were able to backup the hypothesis that location is important factor when analysing SME default by mapping the percentage default by electoral division and local authority in Fig. 5.3.

5.3 Benchmark Models

As mentioned in previous sections, when building predictive models it is essential to have a baseline or benchmark model to compare the experiments to. Therefore benchmark regression models have been built in R and SAS that will be used to make comparisons to experiments in this chapter. Logistic regression was chosen because of its success in experiments in Section 4.2 and its wide use in industry. The models have been trained both using R and SAS using a 70% stratified sample dataset with 30% being kept for holdout which will be used for testing. Results may vary from both R and SAS due to varying samples in each application and algorithms will be slightly different but they should relatively close. This will not impact analysis of comparing experiments as experiments in R will be compared to the benchmark from R and likewise for comparisons
in SAS. The AUC will be the performance measure of choice for evaluating how well the model performs over all possible thresholds. As the AUC does not give the accuracy of the model at a specific threshold recall, specificity and balanced accuracy (BA) will also be assessed. The threshold used is based on fine tuning of the parameters testing using a validation dataset in Section 4.2. The breakdown of the training and test data for the benchmark model and their target class distribution can be shown in Table 5.1.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th># Bad</th>
<th># Good</th>
<th># Observations</th>
<th>Good:Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Delinquency</td>
<td>Training</td>
<td>483</td>
<td>1,565</td>
<td>2,048</td>
<td>76:24</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>245</td>
<td>633</td>
<td>878</td>
<td>72:28</td>
</tr>
<tr>
<td><strong>Previous Delinquency</strong></td>
<td><strong>Total</strong></td>
<td><strong>728</strong></td>
<td><strong>2,198</strong></td>
<td><strong>2,926</strong></td>
<td><strong>75:25</strong></td>
</tr>
<tr>
<td>No Previous Delinquency</td>
<td>Training</td>
<td>474</td>
<td>16,435</td>
<td>16,909</td>
<td>97:03</td>
</tr>
<tr>
<td></td>
<td>Test</td>
<td>177</td>
<td>7,070</td>
<td>7,247</td>
<td>97:03</td>
</tr>
<tr>
<td><strong>No Previous Delinquency</strong></td>
<td><strong>Total</strong></td>
<td><strong>651</strong></td>
<td><strong>23,505</strong></td>
<td><strong>24,156</strong></td>
<td><strong>97:03</strong></td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td><strong>1,379</strong></td>
<td><strong>25,703</strong></td>
<td><strong>27,082</strong></td>
<td><strong>95:05</strong></td>
</tr>
</tbody>
</table>

Table 5.1: Breakdown Holdout Training/Test Dataset for Benchmark Models

The threshold used for performance measuring is based on fine tuning of the parameters testing using a validation dataset in Section 4.2. The results from the benchmark models can be found in Table 5.2 below.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th>Software</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD_Bench_R</td>
<td>Previous Delinquency</td>
<td>R</td>
<td>0.542</td>
<td>0.696</td>
<td>0.619</td>
<td>0.654</td>
</tr>
<tr>
<td>PD_Bench_SAS</td>
<td>Previous Delinquency</td>
<td>SAS</td>
<td>0.529</td>
<td>0.616</td>
<td>0.573</td>
<td>0.615</td>
</tr>
<tr>
<td>NPD_Bench_R</td>
<td>No Previous Delinquency</td>
<td>R</td>
<td>0.525</td>
<td>0.739</td>
<td>0.632</td>
<td>0.671</td>
</tr>
<tr>
<td>NPD_Bench_SAS</td>
<td>No Previous Delinquency</td>
<td>SAS</td>
<td>0.492</td>
<td>0.748</td>
<td>0.620</td>
<td>0.654</td>
</tr>
</tbody>
</table>

Table 5.2: Benchmark Model Results for Experiment Comparison

Models in this this section will be aliased for ease of reading, for example the benchmark models will be referred to as PD_Bench_R, PD_Bench_SAS, NPD_Bench_R and NPD_Bench_SAS which can be seen in Table 5.2 above. The results found here are consistent with the evaluations completed in Chapter 4 which guarantees we have good benchmarks to compare the results in the experiment in this chapter to.

For each of the model results the AUC will be used to measure the the accuracy of the model over all possible thresholds. Performance measures over one threshold will also
be evaluated. Recall, Specificity, Balanced Accuracy will be recorded to evaluate how
the model performs at specified threshold.

5.4 Correlation Analysis

Correlation matrix is a very simple and powerful way of analysing the relationship of
predictive features with each other and the target features. Including highly correlated
predictive features has the potential to throw off or fool your predictive model potentially
causing misleading results. The Pearson correlation coefficient is a common measure
used to test the data for this relationship and will be deployed in this experiment.
The correlation tests are ran separately on each category of features created as part of
this experiment e.g. CSO Features by ED & LA, SME Default Trends by ED & LA,
Homeloan and Personal Loan Default Trends by ED & LA, Transactions Behaviour by
ED & LA, and Binned SME Defaults Rates by ED and LA. Fig. 5.4 shows a subset of
the results demonstrating correlation scores between the experiment features categories
with a full breakdown of the results available in Appendix B where a correlation score
coloured red represents a very strong positive correlation between features and blue a
very low negative correlation.
Chapter 5. *Experiment Implementation & Evaluation*

There is a high level of correlation between many of the macro-economic features generated as part of the experiment. Predictive features with pairwise correlation above 0.80 are considered candidates for removal. For features that were this highly correlated the feature scoring the lowest bivariate correlation to the target feature was removed, ensuring the feature which had the strongest relationship with the target class was prioritised and kept for further predictions. A subset of the results after this feature reduction process can be found in Fig. 5.5 with a full set of results available in Appendix C.

**FIGURE 5.4:** Correlation Analysis of Macro-Economic Features
The most correlated macro-economic features with the target feature are then included in the prediction model and will be evaluated against the benchmark model for performance comparison.
Figure 5.6: Most correlated features with target feature

Table 5.3 details the results for the re-trained model on the previous delinquency training data using the macro-economic features selected as part of correlation analysis.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD.Bench.R</td>
<td>0.542</td>
<td>0.696</td>
<td>0.619</td>
<td>0.654</td>
</tr>
<tr>
<td>PD.Corr5.R</td>
<td>0.534</td>
<td>0.669</td>
<td>0.602</td>
<td>0.654</td>
</tr>
<tr>
<td>PD.Corr10.R</td>
<td>0.556</td>
<td>0.679</td>
<td>0.618</td>
<td>0.658</td>
</tr>
<tr>
<td>PD.Corr15.R</td>
<td>0.594</td>
<td>0.690</td>
<td>0.642</td>
<td>0.665</td>
</tr>
<tr>
<td>PD.Corr20.R</td>
<td>0.561</td>
<td>0.698</td>
<td>0.630</td>
<td>0.665</td>
</tr>
</tbody>
</table>

Table 5.3: Previous Delinquency Model results when most important Macro-economic features calculated using correlation were included in training.

*BA = Balanced Accuracy

The results arising from the inclusion of macro-economic features based on the correlation analysis are compared against the benchmark model (PD.Bench.R) built in Section 5.3. PD.Corr5.R will relate to the model that was trained using the top 5 features from the correlation analysis, PD.Corr10.R is the top 10 features, PD.Corr15.R is the top 15
features, $PD_{Cor20}R$ is the top 20 features. The highest results for each performance metric are highlighted in green.

The results from Table 5.3 are quite promising. Three of the models AUC is better than the benchmark model. $PD_{Cor15}R$ and $PD_{Cor20}R$ in particular perform very strongly across all performance measures. Based on the results $PD_{Cor15}R$ would be selected as the best model as it scored strongly in specificity, balanced accuracy and AUC but also performed exceptionally with a recall of 59.4%, much larger than any of the experimental models and 5.2% higher than the benchmark. The AUC is also 1.1% higher in $PD_{Cor15}R$ compared with the benchmark model.

Table 5.4 details the results for the re-trained model on the no previous delinquency training data using the macro-economic features selected as part of correlation analysis.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>$NPD_{Bench}R$</td>
<td>0.525</td>
<td>0.739</td>
<td>0.632</td>
<td>0.671</td>
</tr>
<tr>
<td>$NPD_{Cor5}R$</td>
<td>0.525</td>
<td><strong>0.743</strong></td>
<td><strong>0.634</strong></td>
<td><strong>0.671</strong></td>
</tr>
<tr>
<td>$NPD_{Cor10}R$</td>
<td>0.555</td>
<td>0.725</td>
<td>0.640</td>
<td>0.677</td>
</tr>
<tr>
<td>$NPD_{Cor15}R$</td>
<td>0.538</td>
<td>0.723</td>
<td>0.630</td>
<td><strong>0.680</strong></td>
</tr>
<tr>
<td>$NPD_{Cor20}R$</td>
<td>0.549</td>
<td>0.731</td>
<td><strong>0.640</strong></td>
<td>0.676</td>
</tr>
</tbody>
</table>

Table 5.4: No Previous Delinquency Model results when most important Macro-economic features calculated using correlation were included in training.

*BA = Balanced Accuracy

The results arising from the inclusion of macro-economic features based on the correlation analysis are compared against the benchmark model ($NPD_{Bench}R$) built in Section 5.3. $NPD_{Cor5}R$ will relate to the model that was trained using the top 5 features from the correlation analysis, $NPD_{Cor10}R$ is the top 10 features, $NPD_{Cor15}R$ is the top 15 features, $NPD_{Cor20}R$ is the top 20 features. The highest results for each performance metric are highlighted in green.

The results from Table 5.4 are also promising. Like the previous experiment three of the models AUC outperform the benchmark model. $NPD_{Cor10}R$ looks the strongest of them all for predicting defaults as it has the largest recall score. $NPD_{Cor15}R$, $NPD_{Cor20}R$ recall and AUC are strong but because $NPD_{Cor10}R$ has less features it is chosen as the best.
5.5 Feature Selection

As mentioned in the research literature in Section 2.9 feature selection is important when building a predictive model for reasons such as reducing the complexity of the model, mitigating the risk of over-fitting, overhead involved with having to understand and maintain a larger number of features, model training and computation time and when you need evaluate/explain your results. In summary, simpler is better in the majority of cases.

Therefore a number of feature selection processes will be experimented with in this section to try and identify the features that are of key importance, reducing the complexity of the model and identifying the important features. The two processes that will be used are Information Gain Feature Importance and Random Forest Feature Importance.

The feature selection process of this experiment was only carried out on the training partitions of the dataset.

5.5.1 Information Gain Feature Importance

Information gain is an approach that utilises measures commonly seen when a decision tree model is being trained (See Section 2.8.4). It calculates and ranks features using entropy and information gain. For each experiment the existing scorecard features and macro-economic features as part of this experiment were analysed. Since we are only interested in the identifying the importance of macro-economic features as part of this research the existing features were stripped out of the result as we cannot tamper with the benchmark model dataset as this could lead to misleading results. Due to the risks previously discussed with having too many features in your training dataset only the top 20 features will be included in models to be trained along with the existing scorecard features as part of this experiment.

Information Gain for Previous Delinquency Data

Addressing the Previous Delinquency dataset the information gain was calculated for each of the existing scorecard features and macro-economic features. Details of the feature importance can be seen in Fig. 5.7.
Figure 5.7: Top 20 Macro-Economic Feature Calculated by Information Gain on Previous Delinquency Dataset

As can be seen in Fig. 5.7 the results of this test are not very promising. The information gain is very small for features which suggests the features were not any better at explaining the target feature than the existing scorecard features.

The features that show strongest performance \textit{ED\_LOAN}, \textit{ED\_HOME} as based on default rates of personal loan and homeloan customers at electoral division. This intuitively could makes sense i.e. if people in an electoral division are struggling with their loan and mortgage repayments they are less likely to spend money in businesses in that area. \textit{ED\_LOWER\_THAN\_UPPER\_SECONDARY}, \textit{ED\_UNEMPLOYMENT} and \textit{ED\_NON\_MANUAL\_OCCUPATION} relates to features created in the census data with low levels of education and high unemployment rate which again intuitively at least make sense. If there are people in an area with low levels of education who cant get work they’re less likely to spend money on businesses in that area.

Despite the low information gain of all these macro economic feature in the results above the model was re-trained and tested. Separate models will be trained using the top 5,
10, 15 and 20 features from the information gain calculation. Table 5.5 details the results for the re-trained model on the test data using the features selected as part of the information gain calculation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD_Bench_R</td>
<td>0.542</td>
<td>0.696</td>
<td>0.619</td>
<td>0.654</td>
</tr>
<tr>
<td>PD_IG5_R</td>
<td>0.540</td>
<td>0.715</td>
<td>0.627</td>
<td>0.652</td>
</tr>
<tr>
<td>PD_IG10_R</td>
<td>0.548</td>
<td>0.693</td>
<td>0.621</td>
<td>0.649</td>
</tr>
<tr>
<td>PD_IG15_R</td>
<td>0.536</td>
<td>0.690</td>
<td>0.613</td>
<td>0.650</td>
</tr>
<tr>
<td>PD_IG20_R</td>
<td>0.544</td>
<td>0.711</td>
<td>0.628</td>
<td>0.651</td>
</tr>
</tbody>
</table>

Table 5.5: Previous Delinquency Model results when most important Macro-economic features calculated using Information Gain were included in training.

*BA = Balanced Accuracy

The results arising from the inclusion of macro-economic features based on the information gain feature importance are compared against the benchmark model (PD_Bench_R) built in Section 5.3. PD_IG5_R will relate to the model that was trained using the top 5 features from the information gain calculation, PD_IG10_R is the top 10 features, PD_IG15_R is the top 15 features, PD_IG20_R is the top 20 features. The highest results for each performance metric are highlighted in green.

None of the models trained in this experiment performed better under the AUC which identifies the best model accuracy of the model over all possible thresholds. This is not really surprising as none of the features exhibited useful information in the information gain analysis. The PD_IG5_R, PD_IG20_R models have a larger specificity than the benchmark meaning the model was able to identify a larger proportion of the negative cases correctly. PD_IG10_R and PD_IG20_R both returned higher balanced accuracy and recall. Higher recall means both models predicted higher percentage of the positive class correctly. Although this is an improvement on the benchmark models the results are not markedly better and would need further investigation to test for statistical significance.

**Information Gain for No Previous Delinquency Data**

Addressing the No Previous Delinquency dataset the information gain was calculated for each of the existing scorecard features and macro-economic features. Details of the
feature importance can be seen in Fig. 5.8

It can be seen in Fig. 5.8 the results of this test are not very promising as with the previous experiment. The information gain is very small for features which suggests the features were not any better at explaining the target feature than the existing scorecard features.

The features that show strongest performance are based on SME default rates in an electoral division (See Table F.1). The features that show strongest performance \textit{ED PERCENT 30 06 2014}, \textit{ED CNT 1 30 06 2014} as based on the default rate of SME customers and the number of SME customers in default at an electoral division. \textit{DIFF 12 2013} to \textit{ED DIFF 06 2012 ED} are based on the percentage change of default rates. If defaulting is trending up in an area this could be indicative of further defaulters, likewise if default rate is going down then consumer spend must be good in these areas and there is smaller risk of default is the future.
Again, despite the low information gain of all macro economic feature in the results above the model was re-trained and tested including the most important features from the information gain calculation. Separate models were trained using the top 5, 10, 15 and 20 features from the information gain calculation. Table 5.6 details the results for the re-trained model on the test data using the features selected as part of the information gain calculation.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPD_Bench_R</td>
<td>0.525</td>
<td>0.739</td>
<td>0.632</td>
<td>0.671</td>
</tr>
<tr>
<td>NPD_IG5_R</td>
<td>0.542</td>
<td>0.736</td>
<td>0.639</td>
<td>0.667</td>
</tr>
<tr>
<td>NPD_IG10_R</td>
<td>0.536</td>
<td>0.736</td>
<td>0.636</td>
<td>0.667</td>
</tr>
<tr>
<td>NPD_IG15_R</td>
<td>0.520</td>
<td>0.732</td>
<td>0.626</td>
<td>0.663</td>
</tr>
<tr>
<td>NPD_IG20_R</td>
<td>0.508</td>
<td>0.730</td>
<td>0.619</td>
<td>0.665</td>
</tr>
</tbody>
</table>

Table 5.6: No Previous Delinquency Model results when most important Macro-economic features calculated using Information Gain were included in training.

*BA = Balanced Accuracy

The results arising from the inclusion of macro-economic features based on the information gain feature importance are compared against the benchmark model \((PD_{Bench\_R})\) built in Section 5.3. \(NPD_{IG5\_R}\) will relate to the model that was trained using the top 5 features from the information gain calculation, \(NPD_{IG10\_R}\) is the top 10 features, \(NPD_{IG15\_R}\) is the top 15 features, \(NPD_{IG20\_R}\) is the top 20 features. The highest results for each performance metric are highlighted in green.

None of four models \((NPD_{IG5\_R}, NPD_{IG10\_R}, NPD_{IG15\_R}, NPD_{IG20\_R})\) outperformed the benchmark model in terms of the AUC which identifies the best model based over all possible thresholds. \(NPD_{IG5\_R}\) performed the strongest by scoring a highest recall and balanced accuracy of models compared. This result is promising as it means this model is capturing almost 2% more of the defaulters. This result could be significant in identifying elevated future default risk among SME customers. Tests would need to be carried out to identify statistical significance.

### 5.5.2 Random Forest Feature Importance

A random forest is a very common and popular method of choosing features for predictive modelling. It is an extension of the previously discussed decision tree (See Section 2.8.4
and Section 2.8.7). The approach is to create large number of decision trees and then combine them together to make a classification, hence why it’s called a forest. For each decision tree in the forest, random subsets of the full training set are used, hence random. Taken together, this is why it is called a random forest. Random forests are also widely used in the process of feature selection as the forest is able to identify the important features of building the model by extracting information from each tree and aggregating it.

For both the Previous Delinquency and No Previous Delinquency experiments in this section 1000 trees will be created to evaluate the importance of all the features. Due to the risks previously discussed with having too many features in the training dataset only the top 20 features will be included in models to be trained along with the existing scorecard features as part of this experiment.

**Feature Importance for Previous Delinquency Data using Random Forests**

Random forest feature importance was calculated for each of the existing scorecard features and macro-economic features in the previous delinquency dataset. Details of the feature importance can be seen in Fig. 5.9 below. Unfortunately, due to the complex nature of the random forest (1,000 trees), it was unclear as to what constituted a good and bad score for variable importance.
It can be seen in Fig. 5.9 that features associated with discretionary and non discretionary spend (See Table A) have strong variable importance after running random forests importance variable. It is worth noting that all of the features selected to have higher variable importance are at an electoral division level opposed to a local authority level. This could signify there has been some shift in customer discretionary and non discretionary spend at the geographic level that is useful for predicting SME defaults or non-defaults. As with the previous experiment calculating information gain (See Fig. 5.8) the SME default rates at an electoral division was also in the top 20 most important features.

Using these results separate models will be trained using the top 5, 10, 15 and 20 features from the random forest feature importance algorithm. Table 5.7 details the results for the re-trained model on the test data using the features selected as part of the random forest feature ranking.
### Table 5.7: Previous Delinquency Model results when most important Macro-economic features calculated using Random Forest feature selection were included in training.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD_Bench_R</td>
<td>0.542</td>
<td>0.696</td>
<td>0.619</td>
<td>0.654</td>
</tr>
<tr>
<td>PD_RF5_R</td>
<td>0.548</td>
<td>0.692</td>
<td>0.620</td>
<td>0.662</td>
</tr>
<tr>
<td>PD_RF10_R</td>
<td>0.556</td>
<td>0.703</td>
<td>0.630</td>
<td>0.655</td>
</tr>
<tr>
<td>PD_RF15_R</td>
<td>0.548</td>
<td>0.698</td>
<td>0.623</td>
<td>0.654</td>
</tr>
<tr>
<td>PD_RF20_R</td>
<td>0.535</td>
<td>0.698</td>
<td>0.617</td>
<td>0.651</td>
</tr>
</tbody>
</table>

*BA = Balanced Accuracy

The results arising from the inclusion of macro-economic features based on the random forest feature selection are compared against the benchmark model (PD_Bench_R) built in Section 5.3. PD_RF5_R will relate to the model that was trained using the top 5 features from the random forest variable importance algorithm, PD_RF10_R is the top 10 features, PD_RF15_R is the top 15 features, PD_RF20_R is the top 20 features. The highest results for each performance metric are highlighted in green.

Results are promising from this experiment. The benchmark model did not outperform the models ran as part of this experiment for any performance measure. PD_RF10_R demonstrated very promising results returning a better result for every performance measure compared to the benchmark model, including a 1.4% increase in classifying customers in default (recall). The only result it was not higher in was the AUC which scored highest in the PD_RF5_R model. Overall PD_RF10_R looks like a promising model and further tests could prove if macro-economic features are significant to this.

**Feature Importance for No Previous Delinquency Data using Random Forests**

Random forest feature importance was calculated for each of the existing scorecard features and macro-economic features in the no previous delinquency dataset. Details of the feature importance can be seen in Fig. 5.10 below.
Figure 5.10: Top 20 Macro-Economic Feature Calculated by Random Forest Feature Importance on No Previous Delinquency Dataset

It can be seen in Fig. 5.10 that features associated with discretionary and non discretionary (See Table A) have strong variable importance after running random forests importance variable. This was also the case in the previous experiment. This could be caused due to spending shift by customers in those electoral divisions going up or down. The random forest variable importance top 20 features are based on electoral division. Lower education rates and occupation rate appear in the top 20 features as they did in information gain experiment in (See Fig. 5.7 and Fig. 5.8).

Using these results separate models will be trained using the top 5, 10, 15 and 20 features from the random forest feature importance algorithm. Table 5.8 details the results for the re-trained model on the test data using the features selected as part of the random forest feature ranking.
Table 5.8: No Previous Delinquency Model results when most important
Macro-economic features calculated using Random Forest feature
selection were included in training.

*BA = Balanced Accuracy

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPD_Bench_R</td>
<td>0.525</td>
<td>0.739</td>
<td>0.632</td>
<td>0.671</td>
</tr>
<tr>
<td>NPD_RF5_R</td>
<td>0.526</td>
<td>0.733</td>
<td>0.630</td>
<td>0.670</td>
</tr>
<tr>
<td>NPD_RF10_R</td>
<td>0.497</td>
<td>0.734</td>
<td>0.615</td>
<td>0.670</td>
</tr>
<tr>
<td>NPD_RF15_R</td>
<td>0.473</td>
<td>0.732</td>
<td>0.602</td>
<td>0.664</td>
</tr>
<tr>
<td>NPD_RF20_R</td>
<td>0.485</td>
<td>0.730</td>
<td>0.608</td>
<td>0.664</td>
</tr>
</tbody>
</table>

The results arising from the inclusion of macro-economic features based on the random forest feature selection are compared against the benchmark model \(PD\_Bench\_R\) built in Section 5.3. \(NPD\_RF5\_R\) will relate to the model that was trained using the top 5 features from the random forest variable importance algorithm, \(NPD\_RF10\_R\) is the top 10 features, \(NPD\_RF15\_R\) is the top 15 features, \(NPD\_RF20\_R\) is the top 20 features. The highest results for each performance metric are highlighted in green.

Results from the macro economic based predictive models are not very promising from this experiment as the benchmark model outperforms the others across all performance measures, apart from \(NPD\_RF5\_R\) which performs slightly better on recall than the benchmark model \(NPD\_Bench\_R\).

### 5.6 Coarse Classification

As mentioned in Section 2.10, coarse classification or binning is a process where you transform features, continuous or nominal into a simplified structure of just a number of categories. There are many benefits which have been noted in the literature for using coarse classification such as handling missing data and outliers and increasing robustness by reducing risks or over-fitting. A standard approach is to split each feature into approximately three to six groups or bins. This is done by finding cut-points in the data and evaluating the relationship with target feature using Weight of Evidence (WoE) and the Information Value (IV) to compare the predictive capability of grouped features.
For both the *Previous Delinquency* and *No Previous Delinquency* dataset experiments in this section coarse classification will applied to all the macro economic features where the aim will be to transform each feature into a much more simplified feature which will have up to 6 groups or bins. Coarse classification will not be applied to existing model features as this would not allow for a fair comparison with the benchmark model. It also worth noting that coarse classification must only be run on the training dataset.

**Coarse classification for Previous Delinquency Data**

Coarse classification will be run on the macro-economic previous delinquency features. As mentioned previously a maximum of 6 bins/groups per feature will be created as per the literature (Hand et al., 2005). As with previous experiments in this chapter the top 5, 10, 15, 20 grouped features will be added to the existing previous delinquency model feature set and models will be generated. The results from these models will then be compared and evaluated against the results from the benchmark model.

Coarse classification is applied to the previous delinquency macro-economic features, results are shown in Fig. 5.11

![Figure 5.11: Feature Ranking by Information Value after Coarse Classification is Applied to the Macro-Economic Features of the Previous Delinquency Dataset](image)
The results from the coarse classification are interesting, there are seven binned features based on local authority and thirteen based on electoral division. In other experiments so far the feature importance has been dominated completely by electoral division features. Most (14) of the binned features are sourced from the discretionary/non-discretionary spend dataset (See Table A). The top rated feature \textit{DIFF PERCENT 06 2012 ED} is the difference in SME default rates at June 2012 and at the observation point June 2014 by electoral division. It is worth noting also that the top five or six binned features appear to be much more predictive based on the information value than the remaining features. This illustrated in in Fig. 5.12.

\begin{figure}
\centering
\includegraphics[width=\textwidth]{figure5_12.png}
\caption{Barplot of Feature Ranking by Information Value after Coarse Classification is Applied to the Macro-Economic Features of the Previous Delinquency Dataset}
\end{figure}

If you look to the left of the barplot above you can see that there are five or six features which have a much higher information value than the rest.

Fig. 5.13 demonstrates the results of the top 5 macro economic binned features.
Fig. 5.13 shows demonstrates how the features \textit{DIFF PERCENT 06 2012 ED}, \textit{ED LIVE LST MTH VS 3MTH AVG LIVE MEAN}, \textit{LA LIVE LST MTH VS 12MTH AVG LIVE MEAN}, \textit{ED DISC LST 3MTH MED VS PRV 3MTH MED SPEND MEDIAN}, \textit{ED LIVE LST MTH VS 6MTH AVG LIVE MEAN} appear in their binned state after coarse classification.

The optimal cut-off for each bin in a feature is calculated using the WoE. Information value is used for ranking binned features importance for prediction.

The group highlighted in the chart above shows the rules for building that bin which is highlighted the equation below

\[
\text{Group 4(Event Rate 37.75)} \quad \text{Values} = 0.05 \leq \text{DIFF PERCENT 06 2012 ED} < 0.16
\]

This equation states that all \textit{DIFF PERCENT 06 2012 ED} values greater or equal to 0.05 and less than 0.16 are to be assigned to bin 4. The event rate details the percentage
of the time this happens. In this example, it occurred 37.75% of the time. It can be observed that all the features in Fig. 5.13 were transformed into features of 5 bins which would have been the optimal result based on WoE and information value.

Using these results, separate models will be trained using the top 5, 10, 15 and 20 features from the coarse classification ranked on information value. Table 5.9 details the results for the re-trained model on the test data using the features selected as part of the coarse classification process.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>PD_Bench_SAS</td>
<td>0.529</td>
<td>0.616</td>
<td>0.573</td>
<td>0.615</td>
</tr>
<tr>
<td>PD_Coarse5_SAS</td>
<td>0.543</td>
<td>0.645</td>
<td>0.594</td>
<td>0.627</td>
</tr>
<tr>
<td>PD_Coarse10_SAS</td>
<td>0.552</td>
<td>0.618</td>
<td>0.585</td>
<td>0.619</td>
</tr>
<tr>
<td>PD_Coarse15_SAS</td>
<td>0.511</td>
<td>0.618</td>
<td>0.564</td>
<td>0.59</td>
</tr>
<tr>
<td>PD_Coarse20_SAS</td>
<td>0.511</td>
<td>0.625</td>
<td>0.568</td>
<td>0.59</td>
</tr>
</tbody>
</table>

Table 5.9: Previous Delinquency Model results when most important Macro-economic features calculated using coarse classification were included. *BA = Balanced Accuracy

The results arising from the inclusion of macro-economic features based on the coarse classification process are compared against the benchmark model (PD_Bench_SAS) built in Section 5.3. PD_Coarse5_SAS will relate to the model that was trained using the top 5 features from the information value calculation, PD_Coarse10_SAS is the top 10 features, PD_Coarse15_SAS is the top 15 features, PD_Coarse20_SAS is the top 20 features. The highest results for each performance metric are highlighted in green.

Results are very promising from this experiment, the PD_Coarse5_SAS and PD_Coarse10_SAS models outperformed the benchmark model (PD_Bench_SAS) over every performance measure suggesting it would be valid to include these macro-economic features in the development of a model in the future. The best model appears to be PD_Coarse5_SAS, it uses a less features than PD_Coarse10_SAS and has a much larger AUC. However PD_Coarse10_SAS has a much larger recall value.

These are strong results, the differences between the benchmark model PD_Coarse5_SAS and coarse classification model PD_Coarse5_SAS performance measures are so large that these macro economic features did improve the model.
Coarse classification for No Previous Delinquency Data

Coarse classification is applied to the no previous delinquency macro-economic features, results are shown in Fig. 5.14

The features chosen by coarse classification are very similar to the previous experiment on the previous delinquency dataset. This time however there are more features based on local authority than on electoral division. This is first time during any of the experiments in this research where local authority features are more dominant than the electoral division features after a feature selection process. The top ranked feature \textit{DIFF PERCENT 06 2013 ED} is the difference in SME default rates at June 2013 and at the observation point June 2014 by electoral division.
In Fig. 5.15 like in the previous experiment there appears to be five strong features ranked by the information value.

Using the results separate models will be trained using the top 5, 10, 15 and 20 features from the coarse classification ranked on information value. Table 5.10 details the results for the re-trained model on the test data using the features selected as part of the coarse classification process.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>NPD_Bench_SAS</td>
<td>0.492</td>
<td>0.748</td>
<td>0.620</td>
<td>0.654</td>
</tr>
<tr>
<td>NPD_Coarse5_SAS</td>
<td>0.492</td>
<td>0.738</td>
<td>0.615</td>
<td>0.664</td>
</tr>
<tr>
<td>NPD_Coarse10_SAS</td>
<td>0.497</td>
<td>0.742</td>
<td>0.619</td>
<td>0.667</td>
</tr>
<tr>
<td>NPD_Coarse15_SAS</td>
<td>0.517</td>
<td>0.741</td>
<td>0.629</td>
<td>0.669</td>
</tr>
<tr>
<td>NPD_Coarse20_SAS</td>
<td>0.500</td>
<td>0.739</td>
<td>0.619</td>
<td>0.665</td>
</tr>
</tbody>
</table>

Table 5.10: No Previous Delinquency Model results when most important Macro-Economic features calculated using coarse classification were included.

*BA = Balanced Accuracy
The results arising from the inclusion of macro-economic features based on the coarse classification process are compared against the benchmark model (£PD Bench SAS) built in Section 5.3. £NPD Coarse5 SAS will relate to the model that was trained using the top 5 features from the information gain calculation, £NPD Coarse10 SAS is the top 10 features, £NPD Coarse15 SAS is the top 15 features, £NPD Coarse20 SAS is the top 20 features. The highest results for each performance metric are highlighted in green.

Results are promising for this experiment but not as promising as the experiment for coarse classification on the previous delinquency dataset. This is because there are not promising results from the models using macro economic features until the introduction of 15 features. Ideally the goal is to keep the number of features in the model small for simplicity and robustness. However the £NPD Coarse15 SAS model has outperformed the benchmark on three of the four performance measures. AUC, recall and balanced accuracy. This means that across all possible thresholds this model is performing better than the benchmark (AUC), by correctly identifying more customers that will go into arrears (recall) and, on average, correctly identifying more customers that will or will not go into arrears (balanced accuracy). This model also introduces a large number of features to training model which could be problematic.

There are two hypothesis suggested as to why this experiment has not been as successful for the no previous delinquency data compared to the previous delinquency data is that. One is that the class imbalance in the no previous delinquency model is so low (97:03) that the information is not getting picked up. The other is that the coarse selection results for previous delinquency data were primarily made up of electoral division features where coarse classification results in this experiment were primarily made up of local authority features. It may be that local authority features are not granular enough.

5.7 Addressing the Imbalance in the Dataset

To address the class imbalance property that exists in the datasets in this experiment a number of methods that manipulate the data by re-sampling have been proposed such as random undersampling of the majority class, random oversampling of the minority class and synthetic sampling of the minority class (See Section 2.11). Re-sampling of the dataset must and will only be carried out on the training set of the data.
It is decided that as part of this experiment random undersampling of the majority class would not be carried out as part of the experiment. This is because there is a possibility of removing important information from the training dataset that would allow the model to identify non-defaulters.

As part of this experiment re-sampling data will be carried by means of oversampling the minority class. These experiments will evaluate whether adding macro-economic features to datasets improved the benchmark results. For practical reasons tests will not be carried out on every macro-economic feature created as part of the research. Instead features will be used from previous experiments where macro-economic demonstrated promising results.

For the previous delinquency dataset macro economic features that were included in the \textit{PD\_RF10\_R} model in Section 5.5.2 were included in the experiment. This model performed the best of all previous delinquency models when macro economic features were included.

For the no previous delinquency dataset macro economic features that were included in the \textit{NPD\_IG5\_R} model in Section 5.5.1 were included in this experiment. This model performed the best of all no previous delinquency models when macro economic features were included.

It is hoped that introducing extra observations of the target class through random oversampling in the training set will improve the accuracy of the models across the performance measures for datasets including macro-economic features

\textbf{5.7.1 Random Oversampling of the Minority Class}

Random oversampling will be carried out using the \textit{ovun.sample} function as part of the \textit{rose} package in R\textsuperscript{\footnote{\url{https://cran.r-project.org/web/packages/ROSE/index.html}}} The \textit{ovun.sample} function generates a balanced sample by oversampling minority class examples. It can also be used to under-sample majority classes or use a combination of over and undersampling to balance the training set.

The results from the oversampling of the previous delinquency and no previous delinquency dataset can be found at Table 5.11 and Table 5.13 respectively.
Oversampling for Previous Delinquency Data

It can be observed in Table 5.11 that after oversampling of the previous delinquency dataset the target class distribution is now 50:50 meaning the dataset is now balanced.

<table>
<thead>
<tr>
<th>Model</th>
<th>Dataset</th>
<th># Bad</th>
<th># Good</th>
<th># Observations</th>
<th>Good:Bad</th>
</tr>
</thead>
<tbody>
<tr>
<td>Previous Delinquency</td>
<td>Training Oversample</td>
<td>1,562</td>
<td>1,552</td>
<td>3,114</td>
<td>50:50</td>
</tr>
<tr>
<td>Test</td>
<td></td>
<td>245</td>
<td>633</td>
<td>878</td>
<td>72:28</td>
</tr>
<tr>
<td>Total</td>
<td></td>
<td>1,807</td>
<td>2,185</td>
<td>3,992</td>
<td>55:45</td>
</tr>
</tbody>
</table>

Table 5.11: Breakdown Holdout Training/Test Dataset for oversampled previous delinquency models

Table 5.12 below details the results from the macro-economic features included in the \(PD_{RF10,R}\) model in Section 5.5.2 retrained on the oversampled previous delinquency balanced training dataset. The retrained model is aliased \(PD_{Over\_RF10,R}\) below. The highest results for each performance metric are highlighted in green.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(PD_{Bench_R})</td>
<td>0.542</td>
<td>0.696</td>
<td>0.619</td>
<td>0.654</td>
</tr>
<tr>
<td>(PD_{RF10,R})</td>
<td>0.556</td>
<td>0.703</td>
<td>0.630</td>
<td>0.655</td>
</tr>
<tr>
<td>(PD_{Over_RF10,R})</td>
<td>0.569</td>
<td>0.662</td>
<td>0.615</td>
<td>0.666</td>
</tr>
</tbody>
</table>

Table 5.12: Oversampled Previous Delinquency Model results when most important Macro-economic features from \(PD_{RF10,R}\) model are included.

*BA = Balanced Accuracy

The results from oversampling are promising. The retrained model \(PD_{Over\_RF10,R}\) 1.1%. This increase appears to be based on oversampling alone and not the introduction of the macro-economic features as there was only a marginal increase in the AUC when \(PD_{RF10,R}\) was tested before. Specificity appears to have performed much worse in the oversampled model but this is most likely due to threshold that was selected is not valid any more and may need to be retrained for the oversampled dataset.

Oversampling for No Previous Delinquency Data

It can be observed in Table 5.13 that after oversampling of the no previous delinquency dataset the target class distribution is now 50:50 meaning the dataset is now balanced.
Table 5.13: Breakdown Holdout Training/Test Dataset for Oversampled No Previous Delinquency Models

Table 5.14 below details the results from the macro-economic features included in the \(NPD_{IG5,R}\) model in Section 5.5.1 retrained on the oversampled previous delinquency balanced training dataset. The retrained model is aliased \(NPD\_Over\_IG5\_R\) below. The highest results for each performance metric are highlighted in green.

<table>
<thead>
<tr>
<th>Model</th>
<th>Recall</th>
<th>Specificity</th>
<th>BA*</th>
<th>AUC</th>
</tr>
</thead>
<tbody>
<tr>
<td>(NPD_Bench_R)</td>
<td>0.525</td>
<td>0.739</td>
<td>0.632</td>
<td>0.671</td>
</tr>
<tr>
<td>(NPD_IG5_R)</td>
<td>0.542</td>
<td>0.736</td>
<td>0.639</td>
<td>0.667</td>
</tr>
<tr>
<td>(NPD_Over_IG5_R)</td>
<td>0.553</td>
<td>0.711</td>
<td>0.632</td>
<td>0.688</td>
</tr>
</tbody>
</table>

Table 5.14: Oversampled No Previous Delinquency Model results when most important Macro-economic features from \(NPD\_IG5\_R\) model were included.

*BA = Balanced Accuracy

As with the previous example the oversampled model has performed very well on the AUC performance measure. This consistency of improvements across datasets is promising. The other performance measures recall, specificity and balanced accuracy did not perform that well but this is most likely caused by using the wrong threshold for the oversampling classification.

It can be confirmed that oversampling should be considered when training a prediction model for the two datasets in this research project.

### 5.8 Interpretation of Results and Experiment Overview

This chapter presented the experiment implementation and evaluation of this project.

An experimental address matching application was designed and built as part of this research project to map 1.4 million personal customers addresses and 28,000 SME addresses to two geographic regions in Ireland known as Electoral Division and Local
Authorities. The experiment was a huge success and 120 macro-economic features were generated and tested as part of this experiment. Based on the success of limited tests work carried out as part of this research AIB are going carry out an enterprise application evaluation where they are going to assess the application built as part of this research against external vendors which provide solutions and software for address matching.

This chapter began with some data exploration where SME default trends were visualised over time. It went onto introduce a geographical informations systems (GIS) application visualised how default appears at local authority level. It was observed that local authorities of cities appear to have a higher percentage of default than rural county areas do. It was also observed that there was a much higher percentage of default on the east coast of Ireland. Also SME default rates by the electoral division in Dublin were analysed and it appears that they are not all part of one homogeneous group. These trends and patterns back up the hypothesis that macro-economic factors have an influence on whether or not an SME defaults on their financial obligation.

Two benchmark models were trained as part of the experiment previous delinquency and no previous delinquency. These were used throughout the experiments to compare to predictive models that included macro-economic features. Feature selection processes such as random forest feature importance and information gain were employed along with correlation analysis to try and identify macro-economic features that would be useful in predicting whether or not a SME customer was likely to default on their financial obligations. These methods did not excel, and were not able to identify clear signs that these features were having any real impact on the model.

Coarse classification was deployed to try and simplify the macro-economic by transforming the features from continuous into grouped features with a number of bins or categories. Feature selection was carried out on these binned features to identify useful features for the prediction model. When building a predictive model including these features on the previous delinquency dataset it was found that these macro-economic features were predictive returning better results than the benchmark model. The same result did not hold true for the no previous delinquency dataset and hypothesis have been proposed for why this is.

There was a large imbalance problem within the target class of the datasets in this experiment. An experiment was carried out that oversampled the minority class for the
previous delinquency and no previous delinquency datasets. It was found in both cases that AUC which measures model performance over all possible thresholds of the model performed better after oversampling had been carried out. Based on this it is suggested that any future models using this data investigate oversampling of the minority class.
Chapter 6

Conclusion

6.1 Introduction

This chapter concludes the thesis and will restate the main findings from the research. The research questions and objectives will be stated again with additional discussion on the contributions to the body of knowledge. An evaluation of the research experiment along with an evaluation of the overall research area will be presented along with any limitations that were observed. Ideas and areas of interest for future work and research will be highlighted followed by the concluding remarks.

6.2 Research Definition and Research Overview

The research carried out as part of this thesis involved reviewing the state of the art literature in the general field of credit scoring with a particular emphasis on the use of macro-economic features as well as available literature on data mining and predictive modelling. This research was used to design and implement experiments and build models to assess the predictive capability of macro-economic features capable of identifying if small medium enterprises (SMEs) are going to default on their financial obligation.

As part of the research the following objectives were achieved:
• Reviewed the relevant literature on credit scoring, the use of macro-economic factors in credit risk modelling, data mining techniques and methods for extracting insights from large datasets and building prediction models

• Designed a scalable and efficient address-matching application to assign addresses in Ireland to geographic regions such as electoral divisions and local authorities

• Generated, collated and identified macro-economic features suitable for predicting if SMEs are likely to default

• Designed and built a benchmark prediction model that could be used to compare experiment results

• Employed feature selection techniques to identify the most predictive macro-economic features such as random forest variable importance, information gain and coarse classification

• Designed an experiment to identify the optimal threshold or cut-off value that’s aim was to maximise balanced accuracy, recall and specificity

• Evaluated a sampling technique to handle the class imbalance issue in the data

6.3 Summary of Contributions to Body of Knowledge and Achievements

The following constitutes the worthwhile contributions to the discipline:

• This research demonstrated how an open source search engine and string distance metrics can be combined to build a scalable and efficient address-matching application.

• Explored SME default rates using a geographical information system (GIS) application on a map of Ireland which supported the hypothesis that macro-economic factors are important when modelling the credit risk of SMEs.
• Demonstrated in this experiment that feature selection techniques, information gain and the random forest importance variable were not successful in identifying macro-economic features that were capable of improving the accuracy of the benchmark model.

• Assessed alternative prediction models to the traditionally used logistic regression for credit scoring and observed that there was no model that significantly outperformed the logistic regression model.

• Demonstrated that when handling imbalanced datasets oversampling can be used to improve the accuracy of your prediction model.

• Demonstrated that after applying coarse classification to macro-economic features they are capable of improving the accuracy of the benchmark model in this research.

6.4 Experimentation, Evaluation, Limitations and Open Problems

The aim of this research project was to generate macro-economic features and assess their capability to predict SME customers that are at risk of defaulting on their financial obligation in the future.

The first objective of this project was to create a mechanism for creating macro-economic features. There was no master addresses database that would allow linking customer addresses to geographic regions so therefore creating this link was the first task. To do this, an address-matching application was built that mapped AIB customer addresses to geographic regions in Ireland called electoral divisions and local authorities. Once there was a way of linking SMEs to local authorities and electoral divisions, macro-economic features could be assigned to that SME. Internal AIB and open data sources were used to generate features in these experiments. The Irish census information was used to generate employment, education and occupation features. Internal data was taken mainly from transaction behavioural data and product default rates such as SMEs, personal loans and homeloans.

Once the macro-economic features had been generated a benchmark predictive model for SME customers could be trained. This model would be used for comparing the results.
of models that were trained using macro economic features. As part of the benchmark building process multiple modelling algorithms were trained and evaluated. It was found that there was no discernible difference between most of the algorithms and that most performed well apart from Support Vector Machines (SVM) polynomial and sigmoid which completely over-fitted.

After the feature generation process was complete there were 116 macro-economic features to be assessed for prediction. Feature selection and reduction techniques were applied. Correlation analysis was performed to reduce features that were linearly correlated and could be redundant. Information gain and random forest variable importance feature selection techniques were applied to identify predictive features. After identifying the most important features based on these feature selection processes, these features were then included in the training model to see if they could improve the performance of the benchmark model. Overall they did not provide any promising results and failed to improve on the modelling accuracy of the benchmark model.

Coarse classification was then applied to the macro-economic features. This is a process of transforming interval and nominal predictive features into grouped features with a small number of categories. It is also a technique widely adopted in credit scoring. After applying this transformation to the macro-economic features the most predictive of these new grouped features were included in the benchmark model. It was found the introduction of these macro-economic features increased the accuracy of the model. This research has proven that macro-economic features can be used to predict SME defaulting and could possibly be applied to other risk areas in AIB also.

Feedback in AIB has been very positive. Geographic data is a data source that the Chief Data Officer wants to increasingly utilise. One of the issues or limitations going forward if these features were to be considered for use in an industry-wide credit scorecard is the potential for the maintenance of these features to be an issue. This data does not exist on the Enterprise Data Warehouse in AIB so it would not be possible to use it until there are more controls and production systems in place.
6.5 Future Work and Research

There are numerous other experiments that could be carried out on this dataset that are worthy of consideration. First of all, features in this experiment were applied only to behavioural credit scoring. Similar experiments could be run on application credit scoring. It could also be included in fields such as marketing.

Statistical tests could be carried out to identify whether there are significant differences between default rates in different regions of Ireland, for example, at a local authority level over a time period.

Macro-economic features could be used for segmentation. This way, locations/areas that are similar to each other based on some features will be grouped together.

Data exploration and visualisation techniques have significant potential as informational tools. GIS applications could be used to illustrate the adverse impact caused by the closure of a significant employer and the related impact on the local economy.

6.6 Conclusion

This chapter concludes the research and experimentation performed in order to assess the predictive capability of macro-economic features in predicting if SME customers will default. It delivered an overview of the research and work that was carried out as well the results that were achieved. Future research topics were introduced whereby the experiment can be continued and improved upon. Some of the risks and limitations with this research were identified.

Overall the experiment has been a success as it showed that macro-economic features could be used to increase the accuracy on the benchmark model. Also the address-matching application created in this research will be tested against vendors who offer the same services to evaluate its full potential.
Appendix A

Discretionary / Non-Discretionary Spend Features

<table>
<thead>
<tr>
<th>Features</th>
<th>Data Type</th>
<th>Feature Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA DISC LST MTH VS PRV MTH SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary spend last month compared to previous month</td>
</tr>
<tr>
<td>LA DISC LST MTH VS 3MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary spend last month compared to mean of the previous 3 months</td>
</tr>
<tr>
<td>LA DISC LST MTH VS 6MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary spend last month compared to mean of the previous 6 months</td>
</tr>
<tr>
<td>LA DISC LST MTH VS 12MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary spend last month compared to mean of the last 12 months</td>
</tr>
<tr>
<td>LA DISC LST 3MTH AVG VS PRV 3MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary mean spend over last 3 months compared to mean of the previous 3 months</td>
</tr>
<tr>
<td>LA DISC LST 3MTH AVG VS 6MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary mean spend over last 3 months compared to mean of the last 6 months</td>
</tr>
<tr>
<td>LA DISC LST 3MTH AVG VS 12MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary mean spend over last 3 months compared to mean of the last 12 months</td>
</tr>
</tbody>
</table>
## Appendix A. Discretionary / Non-Discretionary Spend Features

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Interval Description</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA DISC LST 6MTH AVG VS PRV 6MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary mean spend over last 6 months compared to mean of the previous 6 months.</td>
</tr>
<tr>
<td>LA DISC LST 6MTH AVG VS 12MTH AVG SPEND MEAN</td>
<td>Interval</td>
<td>LA discretionary mean spend over last 6 months compared to mean of the last 12 months.</td>
</tr>
<tr>
<td>LA DISC LST MTH VS 3MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary spend last month compared to median of the previous 3 months.</td>
</tr>
<tr>
<td>LA DISC LST MTH VS 6MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary spend last month compared to median of the previous 6 months.</td>
</tr>
<tr>
<td>LA DISC LST MTH VS 12MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary spend last month compared to median of the previous 12 months.</td>
</tr>
<tr>
<td>LA DISC LST 3MTH MED VS PRV 3MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary median spend over last 3 months compared to median of the previous 3 months.</td>
</tr>
<tr>
<td>LA DISC LST 3MTH MED VS 6MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary median spend over last 3 months compared to median of the last 6 months.</td>
</tr>
<tr>
<td>LA DISC LST 3MTH MED VS 12MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary median spend over last 3 months compared to median of the last 12 months.</td>
</tr>
<tr>
<td>LA DISC LST 6MTH MED VS PRV 6MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary median spend over last 6 months compared to median of the previous 6 months.</td>
</tr>
<tr>
<td>LA DISC LST 6MTH MED VS 12MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>LA discretionary median spend over last 6 months compared to median of the last 12 months.</td>
</tr>
<tr>
<td>LA LIVE LST MTH VS PRV MTH LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary spend last month compared to previous month.</td>
</tr>
<tr>
<td>LA LIVE LST MTH VS 3MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary spend last month compared to mean of the previous 3 months.</td>
</tr>
<tr>
<td>LA LIVE LST MTH VS 6MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary spend last month compared to mean of the previous 6 months.</td>
</tr>
</tbody>
</table>
### Appendix A. Discretionary / Non-Discretionary Spend Features

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Interval</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA LIVE LST MTH VS 12MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary spend last month compared to mean of the last 12 months</td>
</tr>
<tr>
<td>LA LIVE LST 3MTH AVG VS PRV 3MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary mean spend over last 3 months compared to mean of the previous 3 months</td>
</tr>
<tr>
<td>LA LIVE LST 3MTH AVG VS 6MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary mean spend over last 3 months compared to mean of the last 6 months</td>
</tr>
<tr>
<td>LA LIVE LST 3MTH AVG VS 12MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary mean spend over last 3 months compared to mean of the last 12 months</td>
</tr>
<tr>
<td>LA LIVE LST 6MTH AVG VS PRV 6MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary mean spend over last 6 months compared to mean of the previous 6 months</td>
</tr>
<tr>
<td>LA LIVE LST 6MTH AVG VS 12MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>LA non-discretionary mean spend over last 6 months compared to mean of the last 12 months</td>
</tr>
<tr>
<td>LA LIVE LST MTH VS 3MTH MED LIVE MEDIAN</td>
<td>Interval</td>
<td>LA non-discretionary spend last month compared to median of the previous 3 months</td>
</tr>
<tr>
<td>LA LIVE LST MTH VS 6MTH MED LIVE MEDIAN</td>
<td>Interval</td>
<td>LA non-discretionary spend last month compared to median of the previous 6 months</td>
</tr>
<tr>
<td>LA LIVE LST MTH VS 12MTH MED LIVE MEDIAN</td>
<td>Interval</td>
<td>LA non-discretionary spend last month compared to median of the previous 12 months</td>
</tr>
<tr>
<td>LA LIVE LST 3MTH MED VS PRV 3MTH MED LIVE MEDIAN</td>
<td>Interval</td>
<td>LA non-discretionary median spend over last 3 months compared to median of the previous 3 months</td>
</tr>
<tr>
<td>LA LIVE LST 3MTH MED VS 6MTH MED LIVE MEDIAN</td>
<td>Interval</td>
<td>LA non-discretionary median spend over last 3 months compared to median of the last 6 months</td>
</tr>
<tr>
<td>LA LIVE LST 3MTH MED VS 12MTH MED LIVE MEDIAN</td>
<td>Interval</td>
<td>LA non-discretionary median spend over last 3 months compared to median of the last 12 months</td>
</tr>
<tr>
<td>LA LIVE LST 6MTH MED VS PRV 6MTH MED LIVE MEDIAN</td>
<td>Interval</td>
<td>LA non-discretionary median spend over last 6 months compared to median of the previous 6 months</td>
</tr>
<tr>
<td>Feature Description</td>
<td>Interval Description</td>
<td></td>
</tr>
<tr>
<td>-------------------------------------------------------------------------------------</td>
<td>----------------------------------------------------------------------------------------</td>
<td></td>
</tr>
<tr>
<td>LA LIVE LST 6MTH MED VS 12MTH MED LIVE MEDIAN</td>
<td>LA non-discretionary median spend over last 6 months compared to median of the last 12 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST MTH VS PRV MTH SPEND MEAN</td>
<td>ED discretionary spend last month compared to previous month</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST MTH VS 3MTH AVG SPEND MEAN</td>
<td>ED discretionary spend last month compared to mean of the previous 3 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST MTH VS 6MTH AVG SPEND MEAN</td>
<td>ED discretionary spend last month compared to mean of the previous 6 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST MTH VS 12MTH AVG SPEND MEAN</td>
<td>ED discretionary spend last month compared to mean of the last 12 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST 3MTH AVG VS PRV 3MTH AVG SPEND MEAN</td>
<td>ED discretionary mean spend over last 3 months compared to mean of the previous 3 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST 3MTH AVG VS 6MTH AVG SPEND MEAN</td>
<td>ED discretionary mean spend over last 3 months compared to mean of the last 6 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST 3MTH AVG VS 12MTH AVG SPEND MEAN</td>
<td>ED discretionary mean spend over last 3 months compared to mean of the last 12 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST 6MTH AVG VS PRV 6MTH AVG SPEND MEAN</td>
<td>ED discretionary mean spend over last 6 months compared to mean of the previous 6 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST 6MTH AVG VS 12MTH AVG SPEND MEAN</td>
<td>ED discretionary mean spend over last 6 months compared to mean of the last 12 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST MTH VS 3MTH MED SPEND MEDIAN</td>
<td>ED discretionary spend last month compared to median of the previous 3 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST MTH VS 6MTH MED SPEND MEDIAN</td>
<td>ED discretionary spend last month compared to median of the previous 6 months</td>
<td></td>
</tr>
<tr>
<td>ED DISC LST MTH VS 12MTH MED SPEND MEDIAN</td>
<td>ED discretionary spend last month compared to median of the previous 12 months</td>
<td></td>
</tr>
<tr>
<td>Description</td>
<td>Interval</td>
<td>Description</td>
</tr>
<tr>
<td>----------------------------------------------------------------------------</td>
<td>----------</td>
<td>----------------------------------------------------------------------------</td>
</tr>
<tr>
<td>ED DISC LST 3MTH MED VS PRV 3MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>ED discretionary median spend over last 3 months compared to median of the previous 3 months</td>
</tr>
<tr>
<td>ED DISC LST 3MTH MED VS 6MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>ED discretionary median spend over last 3 months compared to median of the last 6 months</td>
</tr>
<tr>
<td>ED DISC LST 3MTH MED VS 12MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>ED discretionary median spend over last 3 months compared to median of the last 12 months</td>
</tr>
<tr>
<td>ED DISC LST 6MTH MED VS PRV 6MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>ED discretionary median spend over last 6 months compared to median of the previous 6 months</td>
</tr>
<tr>
<td>ED DISC LST 6MTH MED VS 12MTH MED SPEND MEDIAN</td>
<td>Interval</td>
<td>ED discretionary median spend over last 6 months compared to median of the last 12 months</td>
</tr>
<tr>
<td>ED LIVE LST MTH VS PRV MTH LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary spend last month compared to previous month</td>
</tr>
<tr>
<td>ED LIVE LST MTH VS 3MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary spend last month compared to mean of the previous 3 months</td>
</tr>
<tr>
<td>ED LIVE LST MTH VS 6MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary spend last month compared to mean of the previous 6 months</td>
</tr>
<tr>
<td>ED LIVE LST MTH VS 12MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary spend last month compared to mean of the last 12 months</td>
</tr>
<tr>
<td>ED LIVE LST 3MTH AVG VS PRV 3MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary mean spend over last 3 months compared to mean of the previous 3 months</td>
</tr>
<tr>
<td>ED LIVE LST 3MTH AVG VS 6MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary mean spend over last 3 months compared to mean of the last 6 months</td>
</tr>
<tr>
<td>ED LIVE LST 3MTH AVG VS 12MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary mean spend over last 3 months compared to mean of the last 12 months</td>
</tr>
<tr>
<td>ED LIVE LST 6MTH AVG VS PRV 6MTH AVG LIVE MEAN</td>
<td>Interval</td>
<td>ED non-discretionary mean spend over last 6 months compared to mean of the previous 6 months</td>
</tr>
</tbody>
</table>
## Appendix A. Discretionary / Non-Discretionary Spend Features

<table>
<thead>
<tr>
<th>Feature Description</th>
<th>Interval Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED LIVE LST 6MTH AVG VS 12MTH AVG LIVE MEAN</td>
<td>ED non-discretionary mean spend over last 6 months compared to mean of the last 12 months</td>
</tr>
<tr>
<td>ED LIVE LST MTH VS 3MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary spend last month compared to median of the previous 3 months</td>
</tr>
<tr>
<td>ED LIVE LST MTH VS 6MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary spend last month compared to median of the previous 6 months</td>
</tr>
<tr>
<td>ED LIVE LST MTH VS 12MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary spend last month compared to median of the previous 12 months</td>
</tr>
<tr>
<td>ED LIVE LST 3MTH MED VS PRV 3MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary median spend over last 3 months compared to median of the previous 3 months</td>
</tr>
<tr>
<td>ED LIVE LST 3MTH MED VS 6MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary median spend over last 3 months compared to median of the last 6 months</td>
</tr>
<tr>
<td>ED LIVE LST 3MTH MED VS 12MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary median spend over last 3 months compared to median of the last 12 months</td>
</tr>
<tr>
<td>ED LIVE LST 6MTH MED VS PRV 6MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary median spend over last 6 months compared to median of the previous 6 months</td>
</tr>
<tr>
<td>ED LIVE LST 6MTH MED VS 12MTH MED LIVE MEDIAN</td>
<td>ED non-discretionary median spend over last 6 months compared to median of the last 12 months</td>
</tr>
</tbody>
</table>

Table A.1: Discretionary Spend and Non-Discretionary Spend features derived for experiment
Appendix B

Correlation Analysis

**Figure B.1:** Grouped Features Correlation Analysis
Figure B.2: SME Arrears Trends Correlation Analysis
Figure B.3: Transaction Visa Debit Correlation Analysis
Appendix C

Correlation Analysis Filtered

Figure C.1: Grouped Features Correlation Analysis after Highly Correlated Features have been Removed
Figure C.2: SME Arrears Trends Correlation Analysis

Figure C.3: Transaction Visa Debit Correlation Analysis after Highly Correlated Features have been Removed
Appendix D

Personal Loans and Homeloans
Default Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED_HOME_RATIO</td>
<td>Interval</td>
<td>Homeloan Default Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_LOAN_RATIO</td>
<td>Interval</td>
<td>Personal Loan Default Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_TOTAL_RATIO</td>
<td>Interval</td>
<td>Homeloan &amp; Personal Loan Default Ratio for each Electoral Division</td>
</tr>
<tr>
<td>LA_HOME_RATIO</td>
<td>Interval</td>
<td>Homeloan Default Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_LOAN_RATIO</td>
<td>Interval</td>
<td>Personal Loan Default Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_TOTAL_RATIO</td>
<td>Interval</td>
<td>Homeloan &amp; Personal Loan Default Ratio for each Local Authority</td>
</tr>
</tbody>
</table>

Table D.1: Personal Customers Arrears Ratio features derived for experiment
### Appendix E

Central Statistics Office (CSO) Features

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>ED_UNEMPLOYMENT_RATIO</td>
<td>Interval</td>
<td>Employment Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_NON_MANUAL_OCCUPATION_RATIO</td>
<td>Interval</td>
<td>Non Manual Occupation Ratio for each Electoral Division</td>
</tr>
<tr>
<td>ED_LOWER_THAN_UPPER_SECONDARY_RATIO</td>
<td>Interval</td>
<td>Lower than Secondary Level/Leaving Certificate Ratio for each Electoral Division</td>
</tr>
<tr>
<td>LA_UNEMPLOYMENT_RATIO</td>
<td>Interval</td>
<td>Employment Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_NON_MANUAL_OCCUPATION_RATIO</td>
<td>Interval</td>
<td>Non Manual Occupation Ratio for each Local Authority</td>
</tr>
<tr>
<td>LA_LOWER_THAN_UPPER_SECONDARY_RATIO</td>
<td>Interval</td>
<td>Lower than Secondary Level/Leaving Certificate Ratio for each Local Authority</td>
</tr>
</tbody>
</table>

Table E.1: (CSO) Features derived for experiment
Appendix F

SME Default Behaviour

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>LA_PERCENT_20/06/2014</td>
<td>Interval</td>
<td>SME Default Ratio for each local authority at June 2014</td>
</tr>
<tr>
<td>ED_PERCENT_20/06/2014</td>
<td>Interval</td>
<td>SME Default Ratio for each electoral division at June 2014</td>
</tr>
<tr>
<td>LA_CNT_20/06/2014</td>
<td>Interval</td>
<td># SME Defaults for each local authority at June 2014</td>
</tr>
<tr>
<td>ED_CNT_20/06/2014</td>
<td>Interval</td>
<td># SME Defaults for each electoral division at June 2014</td>
</tr>
<tr>
<td>DIFF_12_2013_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2013 for each local authority</td>
</tr>
<tr>
<td>DIFF_06_2013_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2013 for each local authority</td>
</tr>
<tr>
<td>DIFF_12_2012_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2012 for each local authority</td>
</tr>
<tr>
<td>DIFF_06_2012_LA</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2012 for each local authority</td>
</tr>
<tr>
<td>DIFF_12_2013_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2013 for each electoral division</td>
</tr>
<tr>
<td>DIFF_06_2013_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2013 for each electoral division</td>
</tr>
<tr>
<td>DIFF_12_2012_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to December 2012 for each electoral division</td>
</tr>
<tr>
<td>DIFF_06_2012_ED</td>
<td>Interval</td>
<td>Difference between number of SME defaults from June 2014 to June 2012 for each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2013_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2013 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2013_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2013 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2012_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2012 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2012_LA</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2012 by each local authority</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2013_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2013 by each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2013_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2013 by each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_12_2012_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and December 2012 by each electoral division</td>
</tr>
<tr>
<td>DIFF_PERCENT_06_2012_ED</td>
<td>Interval</td>
<td>Difference between SME defaults rates from June 2014 and June 2012 by each electoral division</td>
</tr>
</tbody>
</table>

Table F.1: SME default behaviour features derived for experiment
Appendix G

Feature Grouping Based on Home Loans, Personal Loans, SME Loans

<table>
<thead>
<tr>
<th>Feature</th>
<th>Data Type</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>GROUPED_ED_ID_SME_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for electoral division default ratios for SME loans.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_SME_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on electoral division default ratios for SME loans.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_SME_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on local authorities default ratios for SME loans.</td>
</tr>
<tr>
<td>GROUPED_LA_ID_BOTH_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authority default ratios for Home Loans &amp; Personal loans.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_BOTH_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on local authorities default ratios for Home Loans &amp; Personal loans.</td>
</tr>
<tr>
<td>GROUPED_ED_ID_BOTH_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for electoral division default ratios for Home Loans &amp; Personal loans.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_BOTH_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on local authorities default ratios for Home Loans &amp; Personal loans.</td>
</tr>
<tr>
<td>GROUPED_ED_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for electoral division default ratios for Personal loans.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on local authorities default ratios for Personal loans.</td>
</tr>
<tr>
<td>GROUPED_LA_ID_HOMELOANS_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authorities default ratios for Home Loans only.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_HOMELOANS_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on electoral division default ratios for Home Loans only.</td>
</tr>
<tr>
<td>GROUPED_ED_ID_HOMELOANS_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authorities default ratios for Home Loans only.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_ED_ID_HOMELOANS_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on local authorities default ratios for Home Loans only.</td>
</tr>
<tr>
<td>GROUPED_LA_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Ordinal/Nominal</td>
<td>Group/Bin feature based on information gain for local authorities default ratios for Personal loans.</td>
</tr>
<tr>
<td>WEIGHT_OF_EVIDENCE_LA_ID_BRANCH_ADVANCE_ARREARS</td>
<td>Interval</td>
<td>WoE for each bin value based on electoral division default ratios for Personal loans.</td>
</tr>
</tbody>
</table>

Table G.1: Binned/Grouped features based on default ratios of SME, Home and Personal Loans by Electoral Division and Local Authority
Bibliography


URL: [http://core.ac.uk/download/pdf/6363277.pdf](http://core.ac.uk/download/pdf/6363277.pdf)


URL: [http://recipp.ipp.pt/handle/10400.22/136](http://recipp.ipp.pt/handle/10400.22/136)

URL: [http://www.palgrave-journals.com/jors/journal/v60/ns1/abs/jors2008171a.html](http://www.palgrave-journals.com/jors/journal/v60/ns1/abs/jors2008171a.html)

URL: [http://www.bis.org/publ/bcbs128.pdf](http://www.bis.org/publ/bcbs128.pdf)

SMEs around the world: Drivers, obstacles, business models, and lending practices’,

Bijak, K. and Thomas, L. C. (2012), ‘Does segmentation always improve model perfor-
URL: http://www.sciencedirect.com/science/article/pii/S0957417411012243

Bradley, A. P. (1997), ‘The use of the area under the ROC curve in the evaluation of
machine learning algorithms’, Pattern recognition 30(7), 1145–1159.
URL: http://www.sciencedirect.com/science/article/pii/S0031320396001422

URL: http://link.springer.com/article/10.1023/A:1010933404324

Brodersen, K. H., Ong, C. S., Stephan, K. E. and Buhmann, J. M. (2010), The balanced
accuracy and its posterior distribution, in ‘Pattern recognition (ICPR), 2010 20th
URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5597285

Data mining and knowledge discovery 2(2), 121–167.
URL: http://link.springer.com/article/10.1023/A:1009715923555

97(3), 470–487.
URL: http://www.sciencedirect.com/science/article/pii/S0304405X10000413


Carroll, R. J. and Ruppert, D. (1988), Transformation and weighting in regression,
Vol. 30, CRC Press.

learning algorithms, in ‘Proceedings of the 23rd international conference on Machine

URL: http://dl.acm.org/citation.cfm?id=1143865


URL: http://www.jair.org/papers/paper953.html


URL: http://dl.acm.org/citation.cfm?id=1007733


URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4063641


URL: http://journals.cambridge.org/abstract_S0269888910000378


URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=1053964


URL: http://www.sciencedirect.com/science/article/pii/S0933365704001010

Bibliography


URL: http://link.springer.com/article/10.1023/A:1007607513941

URL: http://link.springer.com/chapter/10.1007/11564096_52


URL: http://www.sciencedirect.com/science/article/pii/0304405X86900103

URL: http://www.aaai.org/Papers/KDD/1996/KDD96-014

URL: http://www.sciencedirect.com/science/article/pii/S0377221710006272


URL: http://www.yorku.ca/gisweb/eats4400/boost.pdf


URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6152258

URL: http://dl.acm.org/citation.cfm?id=944968

URL: http://www.sciencedirect.com/science/article/pii/S0304405X06001115

URL: http://imaman.oxfordjournals.org/content/12/2/139.short

URL: http://www.palgrave-journals.com/jors/journal/v56/n9/abs/2601932a.html

URL: http://www.jstor.org/stable/2983268

URL: http://www.sciencedirect.com/science/article/pii/S0957417405000850
URL: http://www.palgrave-journals.com/jors/journal/v61/n10/abs/jors2009129a.html

URL: http://pubs.rsna.org/doi/abs/10.1148/radiology.143.1.7063747

URL: http://pubs.acs.org/doi/pdf/10.1021/ci0342472


URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5128907

URL: http://www.jstor.org/stable/2348414


URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6313426

URL: http://arrow.dit.ie/sciendoc/137/


URL: https://www.researchgate.net/profile/Ron_Kohavi/publication/2352264_A_Study_of_Cross-Validation_and_Bootstrap_for_Accuracy_Estimation_and_Model_Selection/links/02e7e51b71b09e91c000000.pdf


URL: https://www.ifsra.ie/stability/Documents/SME%20Conference/Session%201/Paper%202/Paper.pdf


URL: https://ideas.repec.org/p/cbi/ecolet/03-el-12.html


URL: http://www.tandfonline.com/doi/abs/10.1080/01431160412331331012


URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=4527256


URL: http://adsabs.harvard.edu/abs/1966SPhD...10..707L


URL: http://remotesensing.spiedigitallibrary.org/article.aspx?articleid=1182402

**URL:** [http://works.bepress.com/yliu/12/](http://works.bepress.com/yliu/12/)


**URL:** [http://link.springer.com/chapter/10.1007/1-84628-102-4_3](http://link.springer.com/chapter/10.1007/1-84628-102-4_3)


**URL:** [http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf](http://www.aaai.org/Papers/Workshops/2000/WS-00-05/WS00-05-001.pdf)


URL: http://link.springer.com/article/10.1023/A:1009752403260

URL: http://dl.acm.org/citation.cfm?id=372071

URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5233804

URL: http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=5364367

URL: http://dl.acm.org/citation.cfm?id=584093

**URL:** http://www.sciencedirect.com/science/article/pii/S095741740400096X


**URL:** http://search.proquest.com/openview/4b8a79d5e186d57bfe9536b9af8ed3d/1?pq-origsite=gscholar


**URL:** http://www.aaai.org/ocs/index.php/ICWSM/ICWSM10/paper/viewFile/1441/1852Predicting

**URL:** http://is.muni.cz/do/econ/soubory/aktivity/fai/17809295/FAI_issue2010_01_valova.pdf?so=ex


**URL:** http://doi.acm.org/10.1145/321796.321811


**URL:** http://dl.acm.org/citation.cfm?id=1007734


**URL:** http://www.sciencedirect.com/science/article/pii/S0305054899001495


**URL:** http://ieeexplore.ieee.org/xpls/abs_all.jsp?arnumber=6796856


**URL:** http://eric.ed.gov/?id=ED325505


**URL:** https://aje.oxfordjournals.org/content/159/7/702.full