2014-05-31

Predicting Mortgage Arrears: An Investigation Into the Predictive Capability of Customer Spending Patterns

Jamie Roche
Technological University Dublin

Recommended Citation

Roche, J. Predicting mortgage arrears: An investigation into the predictive capability of customer spending patterns. dissertation submitted in partial fulfilment of the requirements of Technological University Dublin for the degree of M. Sc. in Computing (Data Analytics) May 2014.

Predicting mortgage arrears: An investigation into the predictive capability of customer spending patterns

Jamie Roche

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

May 2014

I certify that this dissertation which I now submit for examination for the award of MSc in Computing (Data Analytics), is entirely my own work and has not been taken from the work of others save and to the extent that such work has been cited and acknowledged within the test of my work.

This dissertation was prepared according to the regulations for postgraduate study of the Dublin Institute of Technology and has not been submitted in whole or part for an award in any other Institute or University.

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

Signed: ________________________________

Date: ________________________________

1 ABSTRACT

The management of credit risk and mortgage arrears has become a very important area in financial services and banking. This dissertation set out to build a statistical model, which incorporates customer spending habits and the current equity value of a property, capable of predicting arrears. Current literature identifies many themes such as negative equity and unemployment that are common occurring factors in mortgage arrears but a multi-faceted approach was required to build a model capable of accurately predicting arrears. Property equity values were included in the model by taking the current outstanding value of the loans and using a property price index to work out the current market value of the property. Transactional data was included in the model as an indicator of the spending habits and trends of the borrowers by deriving monthly values for savings, discretionary spend, necessary living expenses and mortgage payments to give an indication of their overall financial health.

Different modelling techniques were applied to the data along with numerous sampling methods in an endeavour to achieve the best results. The models were evaluated using misclassification costs as well as the more traditional measures such as recall, specificity, precision and overall accuracy. The created models achieved a high level of accuracy in predicting arrears a number of months in advance. Even though much of the existing literature on predictive models for mortgage arrears and default favours the use of Neural Networks for this type of classification, it has been shown here that Decision Trees achieved both the better and most consistent scores. The resulting models created achieved a high level of accuracy and were capable of predicting arrears a number of months in advance of the arrears actually occurring. The overall experiment was a success, and it proved that transactional data and equity values can help to improve the accuracy and predictability of an arrears model.

Key words: Predictive modelling, mortgage arrears, financial crisis, imbalanced datasets

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor, Brendan Tierney, for his help and guidance throughout the dissertation and the many useful pointers and invaluable suggestions.

I would like to thank Johnathan Duggan and Séamus Murphy for facilitating me with access to the necessary data, and providing invaluable support throughout the whole process. Additionally, I would like to thank Kenneth Fox for sharing his extensive insights of the workings and nuances of the data.

My thanks to Dawid for his time and patience in proofreading the document to get it to where it is today. Thanks to Kevin, Tommy, Mike, Fionnuala and Brian for offering suggestions and being available to discuss ideas throughout the project.

Finally, thanks to Louise for her encouragement and support throughout the process as it could not have been completed without her.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1. INTRODUCTION

1.1 Background

The issue of mortgage arrears in Ireland has grown hugely since the onset of the world financial crisis. At the end of September 2013 141,520 Principal Dwelling House (PDH) mortgages in Ireland were in arrears (Central Bank of Ireland, 2013) which represents 18.4% of the total outstanding PDH mortgages in Ireland (768,136). The number of mortgages in arrears has grown at a high rate since 2008 due to a number of factors, but overall figures are only available from September 2009 onwards.

Quarter three 2013 is only the second quarter since quarter three 2009 to show a decline in the total number of arrears (-0.96%) aside from quarter four 2012 (-1.53%). The total number of mortgages in arrears as at the end of quarter three now stands at 18.4% and it appears that the rate of arrears may have plateaued and may start to decrease in the near future.

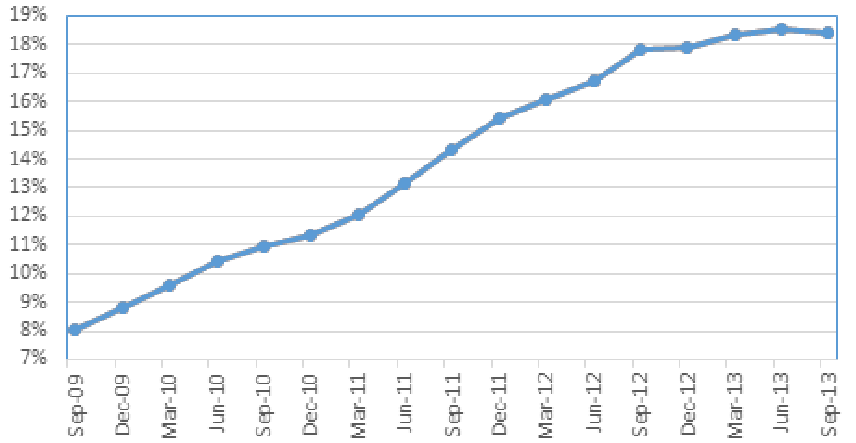


Figure 1.1 Mortgage arrears Sep 2009 to Sep 2013
Source: Data taken from (Central Bank of Ireland, 2013), graph produced by author

While the pace at which mortgage arrears have been increasing has certainly slowed, and indeed the overall level of arrears has decreased for the latest quarter, the mortgage arrears issue is going to go on for quite some time yet. The number of mortgage accounts with arrears of less than 90 days has decreased by 6% quarter on quarter, but the number of accounts with arrears of more than 90 days increased by 1.34%.

There are a number of reasons for a customer to go in to arrears on their mortgage. Traditionally job losses would have been one of the main factors that would have influenced mortgage arrears. However it appears that it is not possible to use an overall or proxy unemployment rate, to better predict the rate of arrears (Gyourko and Tracy, 2014). It would seem rational that this would be the case, given than a national or even local employment rate would not necessarily affect every borrower, or indeed every household. It is also possible that a household could be affected by a loss of employment for one individual from the household, but another individual may be able to cover the payments by themselves.

Many researchers put forward the "Double Trigger" theory as one of the main reasons for default among home owners. (Foote et al., 2008b) explain the Double Trigger Theory as:

> *"This theory holds that default occurs when two things happen simultaneously: the borrower has negative equity and suffers an adverse life event."*

They argue that unemployment in itself is not a sufficient enough indicator of mortgage default. The adverse life event could take the shape of any number of different events, but in many cases this is the loss of a job. If a mortgage holder is in negative equity and has lost their ability to repay the mortgage due to being unemployed, then there is a much higher chance of them defaulting on the mortgage.

In a study of US mortgages it was ascertained that when the negative equity in a mortgage is greater than 50% of the purchase price of the home, half of the defaults associated with these loans are purely as a result of the negative equity (Bhutta et al., 2010). They also found the median borrower will not default until the negative equity of a property reaches -62%. In Ireland and most of Europe, borrowers do not walk away from a mortgage where there is a high level of negative equity, largely due to the fact that the lender will still hold the borrower accountable for any funds not recouped by the sale of the property. In the US there are at least 10 states where non-recourse mortgages are the norm where the property being mortgaged is used as the only collateral on the loan. If the mortgage holder defaults the lender can only recoup any losses against the property, and the mortgage holder is not liable for any outstanding balance.

Typically a non-recourse mortgage will have a Loan to Value (LTV) of 50-60%, so in the case of default the lender should be able to recoup any losses on the loan. However, in some cases the LTV may be higher, but the borrower can still walk away debt-free. Non-recourse mortgages seem to play a part in the strategic defaults of borrowers, though at a certain point it a borrower will likely default even if the lender has recourse against them. Ghent and Kudlyak found that in general, the probability that a borrower will default is roughly 20% higher in states where the lender has no recourse than in states where the lender has recourse over the borrower. The rate of default probability gets higher for mortgages of higher value, with borrowers who have mortgages between $750,000 and $1 million being 66% more likely to default. This shows that recourse can be a very powerful tool for lenders in helping to keep down mortgage default rates.

Strategic defaulters can be split in to two categories: those who strategically default for reasons such as negative equity or high repayments, and those who strategically default as they see it as an opportune time to do so. Those who see it as an opportune time may default as they see many other borrowers defaulting, or they see the banks are in a weakened position. According to (Trautmann and Vlahu, 2013) borrowers are more likely to default on their loans when Banks are in a weakened position or when it is perceived that many other borrowers will default.

There appears to be a social element linked to strategic defaults, whereby if a borrower knows another borrower who has defaulted, then they are more likely to default (Guiso et al., 2013). This does not appear to be a result of defaulters clustering together, but would seem to be that the borrowers know the true cost of default and the uncertainty attached would have become more apparent. The fact that the borrower knows someone else who has defaulted may also have lessened the social stigma.

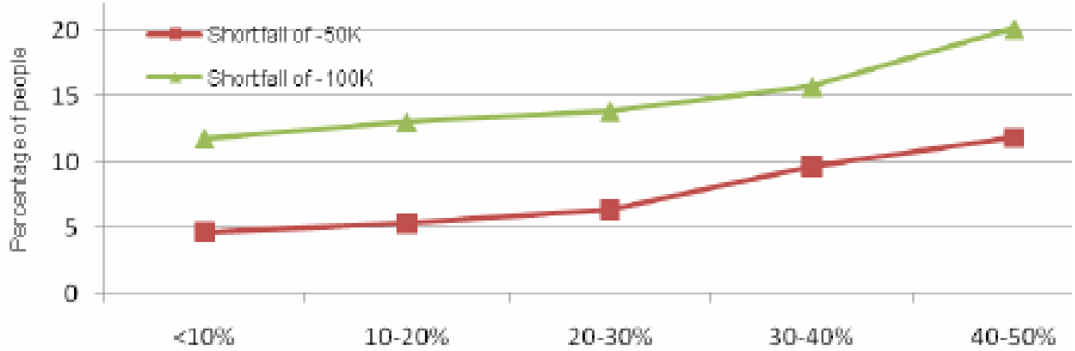


Figure 1.2 Percentage of homeowners willing to strategically default
(Source: Guiso et al., 2013)

According to (Gerardi et al., 2013), numerical ability can help predict the likelihood of a borrower defaulting on a mortgage. They found that borrowers with the worst numerical ability were also the borrowers most likely to default. They reasoned this was not because these borrowers made bad choices when choosing the mortgage type and term, but rather that the individuals were unable to adequately manage their own finances and as a result were more likely to default. Their research may be slightly biased as they focused purely on subprime mortgages in the US, where the level of defaults was the highest amongst borrowers of a lower social demographic. Figure 1.3 shows just how high the default rates were for subprime mortgages taken out between 2003 and 2007, with the general quality of the loans clearly getting worse each year.

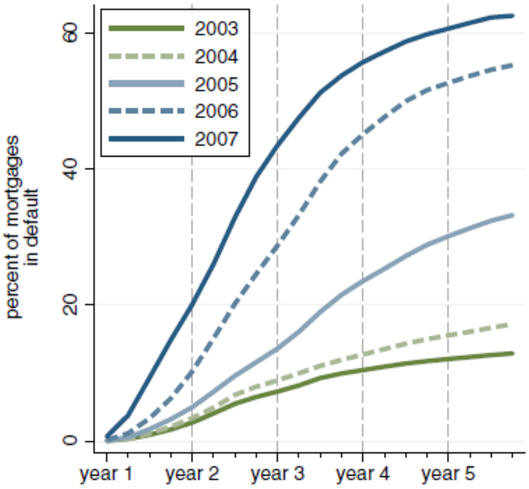


Figure 1.3 Subprime mortgage default rates over 5 year lifecycle, by year of origination
(Source: Gerardi et al., 2013)

It would seem apparent that many borrowers who are at risk of defaulting on their mortgage may be at risk with other credit products they have such as credit cards and personal loans. (Elul et al., 2010) show that borrowers who have high illiquidity rates (defined as an individual with high credit card utilisation) are also likely to be individuals with a high LTV value, who are more likely to default on their mortgage.

The associated costs and perceived decline in social status is seen as a reason for the rate of defaulters, and especially strategic defaulters, being as low as it is. (White, 2010) suggests that the rate of strategic defaulters would be much higher if homeowners in negative equity realised just how much better off they would be if they defaulted. White ascertains that this is partially as a result of individuals not being able to, or wanting to do the complex math to work out the cost of defaulting.

Within the Irish mortgage market (Connor and Flavin, 2013) found that the LTV ratio is one of the main determining factors in mortgages that are likely to default. They found that the LTV or negative equity by itself was not a clear indicator of default but in combination with income stresses, homeowners who had very high LTV rates (equivalent to large negative equity) were more likely to default. This is consistent with the "double trigger" theory, that no one factor by itself is a direct cause of default.

64% of borrowers who took out a mortgage in Ireland between 2005 and 2012 are in negative equity (Duffy et al., 2013). In many cases this is because of the inflated house prices that peaked in 2007. Many of the people who bought in this period were young buyers, who were trying to get on to the property market, but as a result of negative equity are unable to move from the property that they bought.

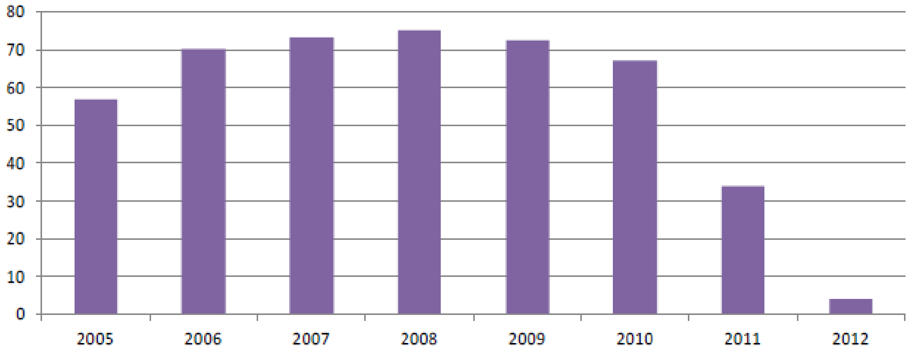


Figure 1.4 Percentage of Irish mortgages in negative equity by year of drawdown
(Source: Duffy, 2013)

(Duffy et al., 2013) also found that the vast majority (63%) of the homeowners who purchased a home between 2005 and 2012 were 40 years old or younger in 2012. These young borrowers are more susceptible to negative equity, with many of these taking out mortgages with high LTV values. These owners are susceptible to any downward trends in house values, as older borrowers would tend to have lower LTV rates on their mortgages. These younger borrowers are also more likely to be at risk of falling in to arrears as their payments are likely to be much higher than older borrowers, as they bought at the height of the property boom.

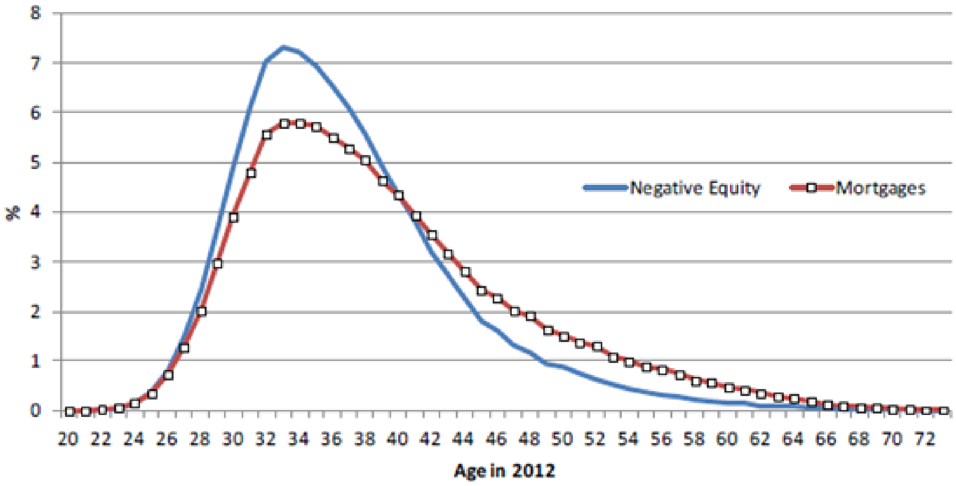


Figure 1.5 Percentage of homeowners in negative equity by borrower age, compared to the total number of mortgages
(Source: Duffy, 2013)

The aim of this brief introduction is to show that mortgage arrears and default has been a very widely studied field, and there are many ways in which mortgages and borrowers can be analysed in order to gain a deeper insight in to why and which borrowers are going to default.

One particular method of investigating mortgage defaults would be by employing data-mining as a tool to extract patterns from the data of lenders, to see what trends exist, and to discover which borrowers are most likely to default. If a lender can successfully predict what borrowers are most at risk of defaulting, then they can take steps to address any issues the borrower may have.

Data Mining is a particularly useful tool to apply to such an exercise, as it can help to discover patterns that are not obvious or discernible without necessarily having a huge amount of subject knowledge. Even if an individual has a great understanding of a certain subject such as mortgage arrears, they will not necessarily be able to tell what borrowers are going to default, and in many cases they will get it wrong. Data Mining should allow techniques to be applied to a dataset to learn more about what is happening in the data, so that the user can garner knowledge from the data.

Data by itself is not of much use, unless you can turn this data in to knowledge. (Fayyad et al., 1996) define Knowledge Discovery in Databases (KDD) as

“*Knowledge Discovery in Databases as the nontrivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data*”.

They outlined their steps from the KDD process which details each of the steps required to be able to gain useful information and knowledge from data held within databases. Their process is nearly 20 years old at this stage, but still holds true for most knowledge discovery projects.

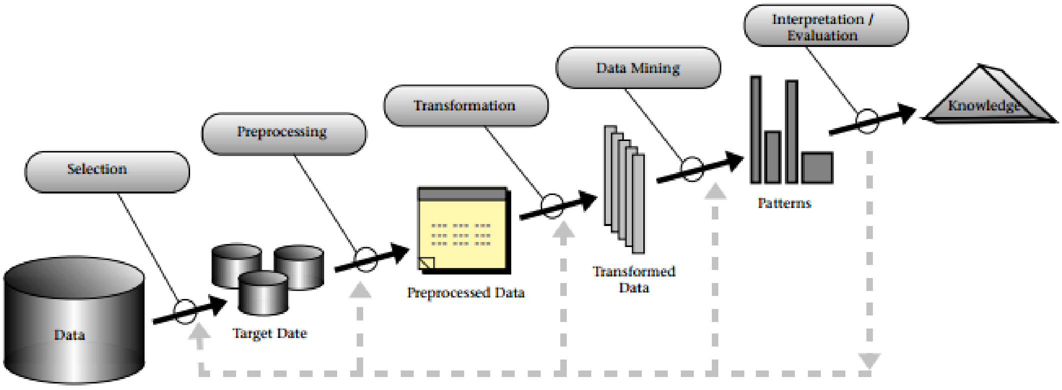


Figure 1.6 Overview of the Knowledge Discovery in Databases (KDD) process
(Source: Fayyad et al., 1996)

It has become common for Data Mining to become synonymous with KDD, but Data Mining is in fact just one part of the KDD process as Figure 1.6 shows. It is however, one of the most important steps as it is the one where the patterns should be identified within the data. Coenen agrees with the fact that Data Mining is just one part of the knowledge extraction process, but to many it has become the term that is used when talking about extracting information from data. Data Mining is in essence a combination of both Machine Learning and Statistics. As a result of this, the field of Data Mining is largely made up of computer scientists and statisticians (Coenen, 2011).

Almost every bank that lends money in the form of mortgages will have models in place to predict both the likely default rate of borrowers who are seeking a mortgage, and to predict default in the existing mortgage book that they hold. KDD and indeed Data Mining plays a large part in these models. These models are generally built on cash flow fundamentals and the borrower’s ability to repay the loan. Much of the data these models use can be quite ‘stale’, as the data may only be refreshed once every few months, or even once a year for some elements of the data. It would appear that credit models existed in most of the Irish banks engaged in mortgage lending during the early 2000s, but anecdotal evidence would suggest that these credit models were largely ignored by banks whose appetite for risk soared in the face of growing profit margins and market share.

1.2 Overview

If a borrower is in difficulty it is expected that the earlier in the process they engage with the lender, the quicker they should be able to get back on track with their payments. Some borrowers may be reluctant to engage with the lender if they are in difficulty as they are scared of the process or are worried it may cause them to lose their home. It would be favourable from the perspective of the lender if they were able to identify customers who are already in difficulty, or are likely to go in to difficulty with their payments. The ability to accurately predict when a borrower is likely to fall behind on their mortgage payments should allow a lender to engage with the borrower earlier, in order to work out a solution to the issue they are having.

This project will investigate the feasibility of taking existing data available on mortgages and borrowers, combined with newly available and derived data on current property values and transaction level data to predict when a borrower might fall into arrears. The borrower transaction data will be broken down in to granular level data as well as aggregated into the broad categories ‘Spend’, ‘Live’, ‘Save’ and ‘Mortgage’. By aggregating the transactions in to these categories it should become apparent if there are changes in a borrower’s behaviour and spending that may indicate their ability to continue paying their mortgage. Particular focus will be placed on the percentage change in the borrower’s credit card balance as well as discretionary spending trends.

At present there are predictive models in place within Lender A that help to model the propensity and probability of borrowers to default, which is defined as being in arrears of ninety days or more on their mortgage. These models are regularly updated and help to reduce the number of people who default by knowing in advance which customers are most likely to be at risk of defaulting. There is presently no model that predicts what customers are likely to go in to arrears. The model proposed by this research will look to fill this gap and predict what customers are likely to miss a payment on their mortgage.

Research on academic literature in both mortgage arrears and predictive models for arrears/default will be carried out throughout the duration of the project. This research will shape the direction of the project to ensure it is relevant to any changes in the literature.

1.3 Aims and Objectives

The primary aim of this project is to assess the predictive capability of customer spend analysis and negative equity estimates, in a mortgage arrears prediction model. The secondary goal is to estimate the optimal point in time for predicting if a customer is going to go in to arrears on their mortgage.

This will be achieved by deriving customer spending habits from transaction level data, and adding this data as an input to an arrears prediction model. Estimates of the current value of each property will also be collected and derived, and will be used to formulate both the relative Negative Equity (NE) and current Loan to Value (LTV) applicable to each property.

The arrears prediction model will be built using predictive modelling techniques with traditional data inputs as well as these new derived data items. The models will be built in the statistical language R using the Rattle data mining interface.

The objectives of the research are:

- To review current literature on mortgage arrears/default in both Ireland and worldwide
- Review research in data mining techniques to find the current algorithms used in arrears/default prediction models
- Undertake the data analysis to identify and derive data required by the model
- Design the experiment to test the hypothesis
- Design and train a prediction model for classifying what mortgages are likely to go in to arrears and ultimately default
- Execute the experiment and compare the predicted results of the model with the actual outcomes

- Analyse results obtained from the models, investigate source of classification errors, and compare results with research carried out
- Assess which model gives the best predictive capability i.e. SVM, Neural Networks, Decision Trees etc.
- Evaluate the success/failure of the experiment
- Determine what future research could be undertaken in the area to expand on the project

1.4 Research Problem

There is an abundance of data available about customers within Lender A, including demographic information, information on products held and transaction level data for every customer. This data is used in a number of ways, including for segmentation, marketing and regulatory reporting. The data available includes details of every transaction carried out by a customer at an individual transaction level. To date this transaction data has not been utilised for segmentation of the customer base, or to try to model behaviour patterns of customers.

The key research problem of this dissertation is to assess if customer transaction level data can be used to improve upon the predictive capability of an arrears prediction model, and to determine if these new inputs will allow the model to predict earlier if a customer is likely to fall in to arrears or not.

1.5 Research Methodology

The research will be split into two main areas. The first area will focus on mortgage arrears and default and what causes borrowers to miss payments on their mortgage and ultimately default on their loan obligations. In this part focus will be put on wider macroeconomic factors to see how the wider financial crisis of recent years has been both a cause and effect of mortgage arrears.

In the second area of the research literature on previous prediction models will be reviewed. Different algorithms and methods will be covered that are used for building prediction models that seek to predict when borrowers are likely to default on their loans.

Thorough understanding of current tools and methods will form the basis of the design for the experiment.

1.6 Scope and limitations

The scope of this dissertation is to build a predictive model within Lender A, which has the capability to accurately predict what mortgages are likely to go in to arrears, before the borrowers actually get in to difficulty. The experiment will use historical data to build and test the model, which will then be validated against previously unseen data.

Even though Lender A has operations outside of the Republic of Ireland, only mortgages from the twenty six counties will be included in the experiment. Buy to Let (BTL) mortgages will be excluded from the model, as they perform in a different way to Primary Dwelling House (PDH) mortgages, which will be the focus of the experiment.

The predictive model that will be built should be able to generalise well, so that it should be possible for the model to be used for new and previously unseen data without major loss of accuracy. A model that generalises well should also negate any possible over-fitting issues within the training dataset.

While the experiment is limited to PDH mortgages within Lender A, theoretically the model produced should be applicable to any mortgage dataset, provided the mortgage data can be structured in the same way. As with any predictive model though, over time the model will need to be re-trained and updated to reflect ongoing changes in the data.

1.7 Organisation of the Dissertation

The remaining chapters of the dissertation are organised in to the Literature Review, Experiment Design, Experiment Implementation, Experiment Evaluation, Conclusions and Future work.

Chapter two is the first of two chapters covering the literature review. This chapter will cover the research carried out on Mortgage Arrears in both Ireland and globally, though there is much more information available for the US mortgage market crash. This chapter will also cover research on macroeconomic factors, and the role they have to play in the mortgage arrears situation at present.

Chapter three will cover the literature review of predictive models – both mortgage arrears prediction models and general prediction models. In this chapter areas of interest in predictive models will be highlighted, along with what algorithms have been used in similar models. The analysis of existing models and techniques to evaluate them presented in this chapter will affect the design of the experiment.

Chapter four will cover the focus of the experiment, its design and implementation. It will also cover how the data is collected and partitioned, as well as the evaluation methods for the models. How the model will be validated against real live data will also be covered in this chapter.

Chapter five will cover the implementation of the models and how the models were designed, trained and tested. It will also cover the different techniques that have been implemented to try to address the class imbalance in the dataset.

Chapter six will be the detailed evaluation of the performance of the model. It will cover in greater detail the results from the model, and how these results can be interpreted in the context of the research carried out. It will cover the evaluation of the performance of the models based on the misclassification costs as well as the other performance metrics outlined such as Recall, Precision and overall Accuracy.

Chapter seven concludes the dissertation and will give an overview of the work carried out. The key results obtained will be reviewed in the context of the literature review, and the contributions to the body of knowledge will be presented. Future areas for research and experimentation will be discussed.

2. LITERATURE REVIEW MORTGAGES

2.1 Introduction

This chapter will cover current research literature in the field of general lending arrears, but will focus primarily on mortgages arrears. Research from around the world in the field of arrears will be covered, though the emphasis will be on the Irish mortgage market. Since the start of the global financial crisis in 2008 there has been a lot of focus on the US mortgage market, with the onset of the failure of sub-prime mortgages an area of intense focus within academic circles. Much of the literature available on mortgage arrears/default is based on US mortgages, but some of the same theory is applicable to the Irish mortgage market.

The research will initially look at macroeconomic factors that have had an influence on the mortgage market in Ireland, as well as the subsequent decrease in housing values and the current arrears situation. The chapter will then focus on specific topics and reasons for the rise of mortgage arrears, and will finish by covering how lenders have been trying to address the arrears situation.

2.2 Macroeconomic factors affecting mortgages

The financial crisis that has persisted around the world since 2008 brought to an end a never seen before period of growth in the Irish economy. The economy had been growing on a near exponential growth rate since coming out of the last recession in the 1980s, and suffered a sharp and painful contraction. To date it has been struggling to recover, with austerity measures implemented by the government helping to improve the fiscal situation, but this has resulted in the economy stagnating.

In the 1980s disposable income levels were low due to unemployment which averaged around 14% along with little economic growth (McQuinn and O’Reilly, 2008), and high levels of personal taxation with over 40% of Irish taxpayers paying tax at a marginal rate of 45% or more (O’Toole, 1993). The “Celtic Tiger” as it became known, brought huge investment in the country with the country reaching almost full employment levels. There are many factors that helped the growth seen in the economy, not least of which was Ireland joining the European Monetary Union in the late 1990s. But perhaps the single biggest driver of the economy was the construction trade. The level of housing units being completed in Ireland was far and above the rates of any other country in the EU, with the exception of possibly Spain. Figure 2.1 details the number of housing units completed in 2007 per 1,000 of population.

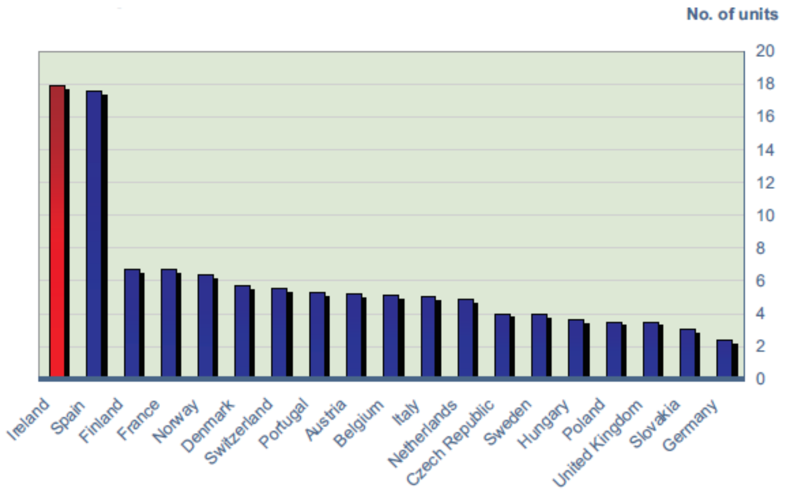


Figure 2.1 Housing unit completions per 1,000 of population in 2007
Source: (Central Statistics Office, 2008)

The average rate across the EU was 5.3 units, but Ireland in 2007 had a completion rate of 18 units per 1,000 of population. For comparison the annual average number of residential units being completed in Ireland during the period 2004 to 2006 was 85,000, while the figure for the UK standing at just over 200,000 (McCarthy and McQuinn, 2013). This is all the more staggering looking at the population figures for 2005 with Ireland having 4.1 million, and the UK having nearly 15 times this at 60 million (Eurostat — Statistical Office of the European Communities, 2007).

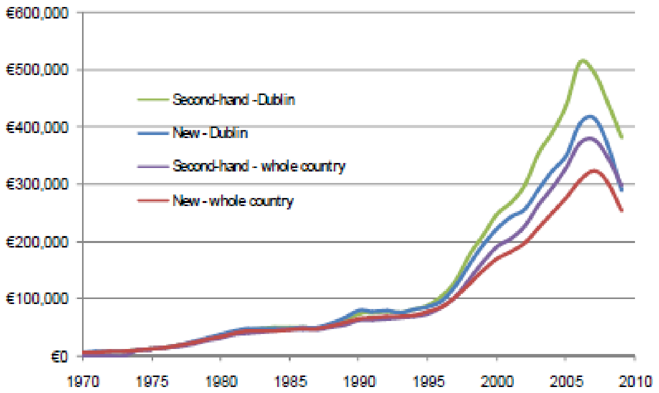


Figure 2.2 Irish house prices 1970 – 2009
Source: (Connor et al., 2012)

The massive increase in the number of houses being built was as a direct result of the “Celtic Tiger” with more and more people able to afford their own homes, and many of these wished to upgrade to better quality housing. The demand for housing units far outstripped the supply, which had the effect of pushing up prices. After a period of little or no growth in the 1980s and early 1990s, Irish property prices grew at an ever increasing speed from roughly 1997 before peaking in 2007 as detailed in Figure 2.2.

The construction industry was one of the largest contributors to the Irish economy throughout the economic boom, with the stamp duty tax returns for 2006 accounting for approximately 17% of all tax returns (Connor et al., 2012). Between 1980 and 1990 the tax returns for property ranged between 4 and 5.3% of all returns (O'Toole, 1993). With other revenues from construction such as income tax on wages and VAT on property sales, the construction boom was the major factor in the huge economic growth seen by the country. The rate of employment in construction grew from just under 7% in 1990 to over 13% in 2007, after which there was a sharp contraction in 2008.

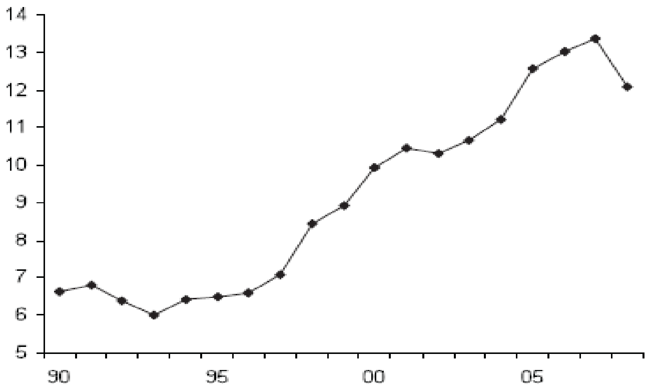


Figure 2.3 Employment in construction from 1990 – 2008 as a percentage of total employment
Source: (Honohan, 2009)

2.2.1 Access to new sources of cheap credit

Irish lending institutions changed their funding sources from being almost fully funded by customer deposits in 1997, to using short term interbank lending and international bond issues up to and including 2008 (Lane, 2011) ,(McCarthy and McQuinn, 2013) and (Kelly, 2009). From 1997 onwards, a large divergence started in the gap between the deposits held by Irish institutions and the loans they had sold as seen in Figure 2.4. This had the undesired effect of increasing the loan to deposit ratio of banks, though at the time this did not seem to deter their lending activity. Traditionally banks governed their own lending levels, and ensured that they were not above a certain level with the loan to deposit ratio, as this was seen as a measure of the risk the bank was exposed to.

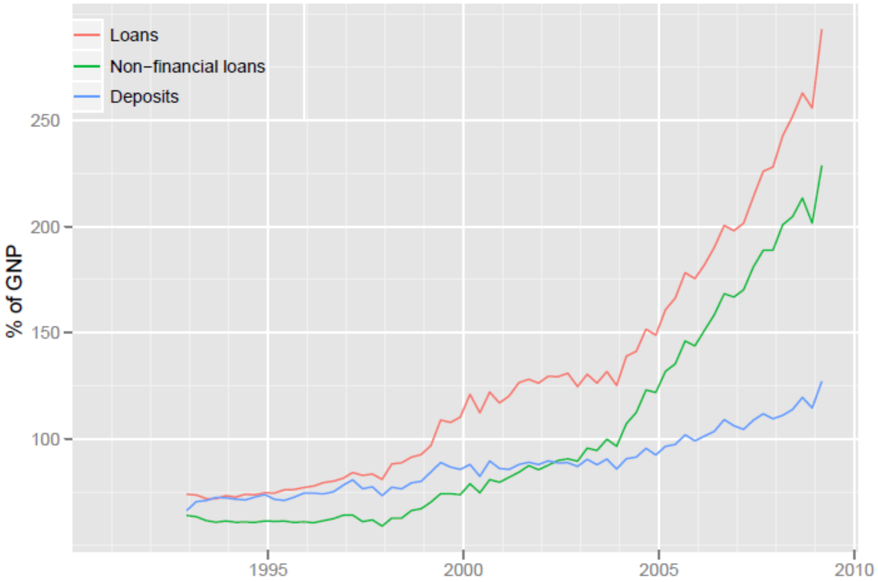


Figure 2.4 Loan to deposit ratios 1993 - 2009
Source: (Kelly, 2009)

Irish banks fuelled their lending through wholesale markets with cheap interbank loans, and this facilitated the advancing of cheap loans to both prospective homeowners and speculative property developers. The Irish mortgage market in 1997 was €20Bn (in 2009 prices) with the lending to property developers standing at €10Bn. By 2008 the figure for mortgage lending was closer to €140Bn, and the amount advanced to property developers was roughly €110Bn (Kelly, 2009). The freely available funding the Irish banks were using dried up in September 2008, when Anglo Irish Bank were reportedly unable to renew their short-term funding in the wholesale markets (Kelly, 2009) & (Honohan, 2009). At the time the other main Irish banks appeared to have little or no difficulty with their funding, but the government feared there would be a risk of contagion if they did not step in to act. The funding in the wholesale markets dried up as banks were much more reluctant to lend to other banks in the wake of the collapse of Lehman Brothers (Lane, 2011), so the Irish banks were unable to re-finance their lending, or at least not at levels that were sustainable. Up until the point that Anglo Irish Bank were unable to renew their funding in the wholesale markets, the Irish Government were adamant that the Banks were in good health, but their actions at the end of September belied this (Kelly, 2009). They felt that if they didn’t act then the whole Irish banking sector would struggle. To this end they implemented the blanket bank guarantee scheme, which covered all deposits and liabilities at each of the covered institutions.

2.2.2 Easing of lending standards during boom

During the boom years of speculative lending to property developers and rising mortgage lending, one of the main banks to the fore of the credit bubble was Anglo Irish Bank. According to (Honohan, 2009), a key metric that is used by regulators to identify banks exposed to increased risk is the rapid expansion of their balance sheet. Any growth of 20% or more is seen as a sign that the balance sheet is growing too quickly and may lead to problems. All of the Irish banks had at least one year of this rate of balance sheet growth, while Anglo Irish Bank had 20% or more balance sheet growth in eight out of nine years during the boom and an average balance sheet growth from 1998-2007 of 36%. This should have set alarm bells ringing in the offices of the Financial Regulator (Honohan, 2009) and (Connor et al., 2012). Irish Nationwide – another rogue lender during the credit bubble, averaged over 20% growth in their balance sheet during the same period and actually surpassed the 20% mark in six of the nine years.

The other domestic banks were under pressure from the market to increase their own lending to catch up to Anglo Irish, who were seen as the pace-setters at the time. To gain further market share in both the domestic mortgage market and in developer’s loans, the other institutions were forced to relax their own lending standards (Honohan, 2009). In combination with the lower pricing of funding this lead the banks to start lending higher levels of funds with higher LTV values (see Figure 2.5) and longer terms than would have previously been unacceptable.

Irish Nationwide was a building society setup as an institution to provide mortgages for its members, but when it saw the money to be made from developer loans it followed the banks, and soon 80% of its lending was to property developers (Connor et al., 2012). The financial regulator through its inaction to prohibit this activity was a willing participant in the credit bubble.

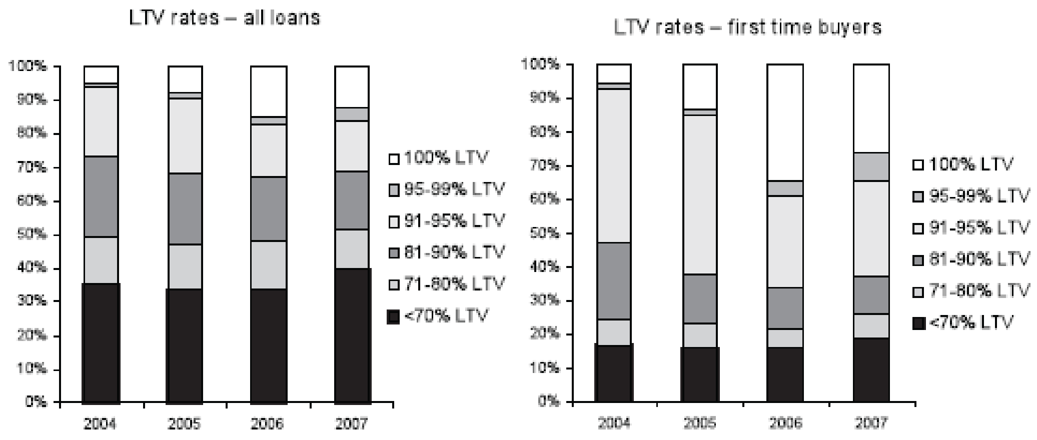


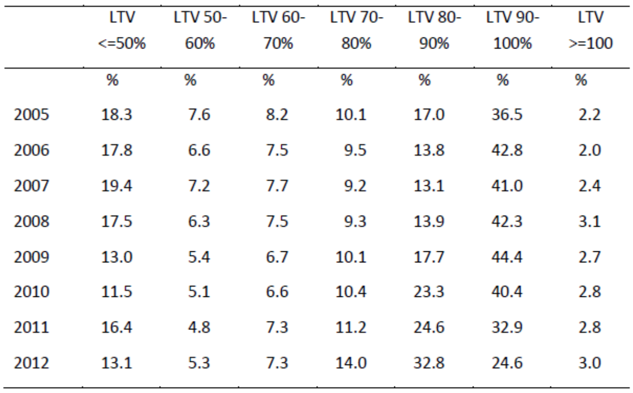
Figure 2.5 Loan to Value rates for mortgages 2004-2007
Source: (Honohan, 2009)

(Honohan, 2009) explains that historically in Ireland Building Societies would have had the lion's share of mortgage lending, and up until the 1980s the traditional banks only had a small share of mortgages. Banks had not been exposed to the problems previously linked with mortgages, in times of financial turmoil and were naïve with their lending practices and ill-positioned to deal with the downturn. (Honohan, 2009) also argues that without the access to the wholesale financial markets, the property boom could never have sustained, as the banks would not have been able to get access to the funds required.

The banks introduced new products during the credit bubble and variations on regular mortgages, which caused numerous issues. Tracker mortgages were introduced, that allowed a borrower to pay interest on their mortgage at a fixed (typically 1%) level above the ECB rate. The banks introduced these products at a time when ECB rates were relatively high, and the cost of funding the mortgages was sustainable. The ECB rate was 2.75% in June 2006 (RTÉ.ie Business news, 2013), so this meant that the tracker rate for most mortgages would be approximately 3.75%. Since the fall of the ECB rates these mortgages have become a burden on the banks, as the cost of funding far outweighs the interest being charged on the loans. The current ECB rate sits at 0.25% (RTÉ.ie Business news, 2013), so tracker rates on average would be 1.25%. The cost of funding a tracker mortgage is considerably more than this, so banks are making a substantial loss on their tracker mortgages. Some of the banks try to offset these losses by increasing their variable rates to compensate for the difference (Goggin et al., 2012).

Table 2.1 Distribution of Loan to Value rates by year of drawdown

Source: (Duffy et al., 2013)

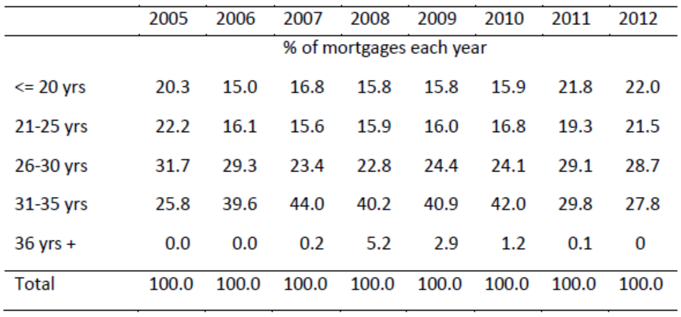


	LTV <=50%	LTV 50- 60%	LTV 60- 70%	LTV 70- 80%	LTV 80- 90%	LTV 90- 100%	LTV >=100
	%	%	%	%	%	%	%
2005	18.3	7.6	8.2	10.1	17.0	36.5	2.2
2006	17.8	6.6	7.5	9.5	13.8	42.8	2.0
2007	19.4	7.2	7.7	9.2	13.1	41.0	2.4
2008	17.5	6.3	7.5	9.3	13.9	42.3	3.1
2009	13.0	5.4	6.7	10.1	17.7	44.4	2.7
2010	11.5	5.1	6.6	10.4	23.3	40.4	2.8
2011	16.4	4.8	7.3	11.2	24.6	32.9	2.8
2012	13.1	5.3	7.3	14.0	32.8	24.6	3.0

Historically in order to get a mortgage, the minimum down payment would have been 8% of the overall purchase price, but in many cases the payment would have been up to 20%. In the midst of the property bubble in Ireland, many of the Irish banks introduced mortgages with LTV rates of up to 100%. This meant that a prospective borrower could borrow 100% of the costs required to purchase a new property. According to (Duffy et al., 2013) the median LTV rose from 85% in 2005 to 89% in 2006, and LTV values of 90% persisted in to 2009 and 2010, even though property prices at that stage had been falling for a number of years. Table 2.1 details the LTV rate for mortgages drawn down between 2005 and 2012, and shows the increasing levels of mortgages that were taken out with LTV values above 80% throughout.

Table 2.2 Initial mortgage term by year of drawdown, % of mortgages each year

Source: (Duffy et al., 2013)



	2005	2006	2007	2008	2009	2010	2011	2012
			% of mortgages each year					
<= 20 yrs	20.3	15.0	16.8	15.8	15.8	15.9	21.8	22.0
21-25 yrs	22.2	16.1	15.6	15.9	16.0	16.8	19.3	21.5
26-30 yrs	31.7	29.3	23.4	22.8	24.4	24.1	29.1	28.7
31-35 yrs	25.8	39.6	44.0	40.2	40.9	42.0	29.8	27.8
36 yrs +	0.0	0.0	0.2	5.2	2.9	1.2	0.1	0
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

Coupled with mortgage LTV values that were very high, the Irish Banks were lending mortgages to borrowers with increasingly longer terms. Prior to the onset of the property bubble the average term length of a mortgage was 20 years, but banks were suddenly willing to lend to borrowers for 25 years and more during the property bubble. Table 2.2 details the initial mortgage term broken down by year of drawdown, from 2005 to 2012. The terms of mortgages shifted towards longer terms, with some mortgages being approved for terms over 35 years.

According to (Kelly, 2009), the average first time buyer in 1995 needed to take out a mortgage equal to three year's average earnings, with the average house costing 4 year's average earnings. In 2006 the average mortgage taken out by a first time buyer was 8 times average earnings, with the average new house costing 10 times average earnings. Figure 2.6 shows the rise of mortgage debt in Ireland, and how it overtook net wealth for the first time in 2007, and in 2008 far surpassed it.

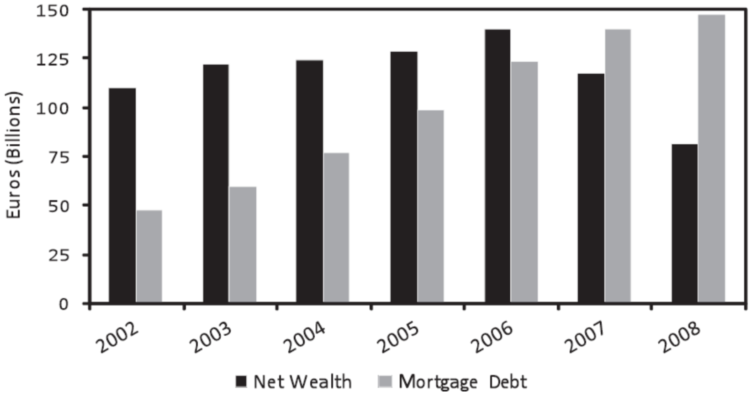


Figure 2.6 Household Net Wealth & Mortgage Debt 2002 – 2008
Source: (McCarthy and McQuinn, 2011)

2.3 The decline of Mortgage Credit Quality

The Irish economy was not the only one to experience a boom in the recent past, as the US experienced its own property lending spree, but it was of a different nature to that of the Irish one. Historically, US home ownership rates were approximately 64% for the two decades leading up to 2004, but by 2007 the rate had increased to 69% (Connor et al., 2012). This large jump in the number of people who owned a home was due in large part to increases in subprime mortgage lending. Subprime mortgage lending is the lending of mortgages to borrowers who would normally be deemed too risky to lend to. This increase in subprime lending can be attributed back to political aspirations, when the Bush administration pushed to increase home ownership among Americans who were not traditional homeowners (Connor et al., 2012). As part of this strategy, the Federal National Mortgage Association (commonly known as Fannie Mae), announced it was going to purchase $2 trillion in home loans for poor and minority households during the period 2000 to 2010 (Connor et al., 2012). This was a very ambitious project, and ultimately caused many of the recent problems with arrears and defaults in the US.

There are many researchers who partially or fully attribute the global financial crisis to the performance of the US mortgage market. (Ellis, 2008a) directly links the global financial crisis, to the losses sustained on US mortgages. She attributes much of the performance of these loans to the reduced lending standards that were applied to mortgages in the US property boom. Some of the causes of this are the reduced quality or availability of documentation on a mortgage (and specifically earnings and income data), higher LTV amounts being lent to borrowers and loans where there was little or no principal paid down in the early years of the mortgage.

The US crisis was different from most of the other crises around the world, not least by the fact that it was the one that triggered the worldwide funding shortage. In Ireland and much of the rest of Europe when a borrower lends money to a borrower for a mortgage, the lender holds the debt on their balance sheet. In the US the general model was to give the loan to the borrower, but then sell the loan on to another institution who would securitise the loans and package them up in to mortgage securities before re-selling them on to another institution. This US model is referred to as the “originate-and-distribute” model, whereas when the lender holds on to the debt on their balance sheet it is called the “originate-and-hold” model (Connor et al., 2012).

These Mortgage Backed Securities (MBS) along with the increased lending to subprime borrowers is seen as one of the main catalysts for the US mortgage crisis. These MBSs caused many problems in the US as the institution holding an individual mortgage as part of a MBS, was far removed from the lender with whom the loan originated. The lender with whom the loan originated had sold on the debt, and had very little reason to monitor the performance or quality of the loan (Connor et al., 2012).

In the US charities were set up to help prospective borrowers with down payments, or deposits on houses they wished to purchase. In many cases these charities were funded by mortgage vendors, with the end goal being able to lend more money to borrowers. If a borrower was able to secure a down payment for a property from one of these charities where they did not raise the funds for the payment from their own finances, then the indication was that the credit quality of the resulting mortgage would be less than that of a mortgage where the borrower saved the money for the down payment (Ellis, 2008a). This assistance for home-buyers also had the result of effectively raising the actual LTV values for these mortgages, as the borrower would have provided no monetary input in the purchase of the house and the LTV for the property would have in fact been 100%.

In Ireland there were no down payment charities that assisted with a deposit for a property purchase, though it was common practice that a family member or friend would help with the deposit for a house if the borrower(s) were unable to finance it themselves – especially for first time buyers. It was also common practice for parents of first time borrowers to act as a guarantor for the mortgage, and this had the effect of allowing first time buyers to borrow higher level of funds. Guarantors are obligated to co-sign a credit agreement along with the main borrower, but they are very rarely required to step in and intervene (Dufhues et al., 2011).

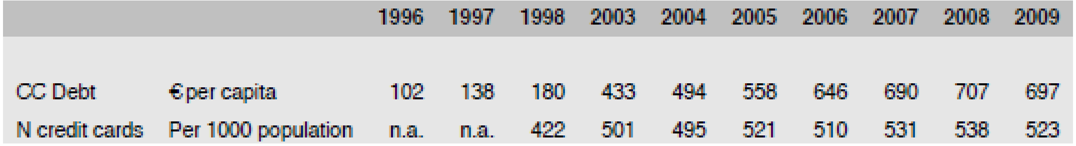
2.3.1 Household indebtedness

Much of the research has focused on both financial and numerical literacy to explain why certain people struggle with their finances. (Klapper et al., 2013) found that people in Russia with poor financial literacy are more likely to have credit portfolios with high APRs and are more likely to experience income shocks. They are also more likely to make poor financial decisions and to engage in short term lending at a greater cost of borrowing. Russia is different to most markets though, as there was little or no personal credit there before 2001 (Klapper et al., 2013). (Gathergood, 2012) found that both self-control and financial literacy were positively correlated to over-indebtedness. He argues that increased financial knowledge in itself will not stop people from getting in to positions of over-indebtedness, as the impulsive behaviour of individuals will cause them to continue to make rash decisions, and use sources of credit at high rates.

(Gerardi et al., 2013) notes that individuals who have poor numerical ability are more likely to get in to difficulty with mortgage repayments, and ultimately get in to arrears or default on a mortgage. They reason that an individual with low numerical ability may not understand the complexities of a product such as an Adjustable Rate Mortgage (ARM), where the borrower has an initial low rate of payments but after a number of years the payments increase substantially.

Mortgage lending was not the only area of credit lending to grow during the Celtic Tiger era and the subsequent credit bubble up until 2008. Table 2.3 shows the per capita credit card debt for Ireland from 1996 to 2009, and there is a marked increase with the average credit card per person in the country standing at €707 in 2008 (Russell et al., 2011). This represents an increase of nearly 700% since 1996.

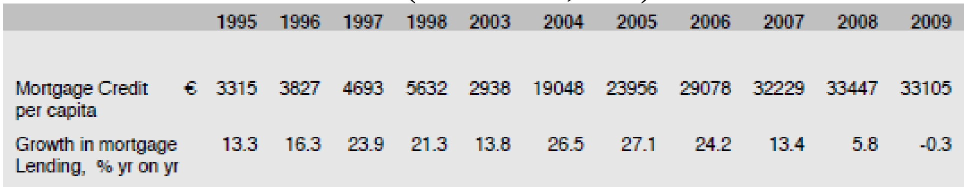
Table 2.3 Credit Card Debt in Ireland 1996-2008
Source: (Russell et al., 2011)



		1996	1997	1998	2003	2004	2005	2006	2007	2008	2009
CC Debt	€per capita	102	138	180	433	494	558	646	690	707	697
N credit cards	Per 1000 population	n.a.	n.a.	422	501	495	521	510	531	538	523

In Table 2.4 the increase in per capita mortgage credit as well as the average credit amount is documented over the period 1995 to 2009. Again the per capita amount peaks in 2008 at €33,447, which is a growth of 1000% from the 1995 figure (Russell et al., 2011).

Table 2.4 Residential Mortgage Credit in Ireland 1995-2008
Source: (Russell et al., 2011)



		1995	1996	1997	1998	2003	2004	2005	2006	2007	2008	2009
Mortgage Credit per capita	€	3315	3827	4693	5632	2938	19048	23956	29078	32229	33447	33105
Growth in mortgage Lending, % yr on yr		13.3	16.3	23.9	21.3	13.8	26.5	27.1	24.2	13.4	5.8	-0.3

It is clear from both the figures on credit card debt and mortgage credit that the average level of debt held by Irish citizens has increased dramatically since the middle of the 1990s. With the onset of the financial crisis however, the average disposable income figure has fallen 8.1% since its height in 2007 (Kelly, 2012). As a result Irish households are having to sustain high levels of debt with smaller incomes. With unemployment rising from 4.5% to 12% in just two years (McCarthy and McQuinn, 2011), certain sectors such as construction have been badly hit, and households are struggling to make the payments on credit they took out in the boom years.

2.3.2 Increase in Mortgage Arrears

The number of PDH mortgages in arrears at the end of quarter three 2009 was 63,619 or 8.01% of the total outstanding mortgages (Central Bank of Ireland, 2013). The rate of arrears increased to 18.54% at the end of quarter two 2013, but achieved its first decrease during quarter three 2013, dropping to 18.42%. However, the rate of mortgages in arrears of over 90 days actually increased from quarter two 2013, with 12.9% of all mortgages in arrears of over 90 days.

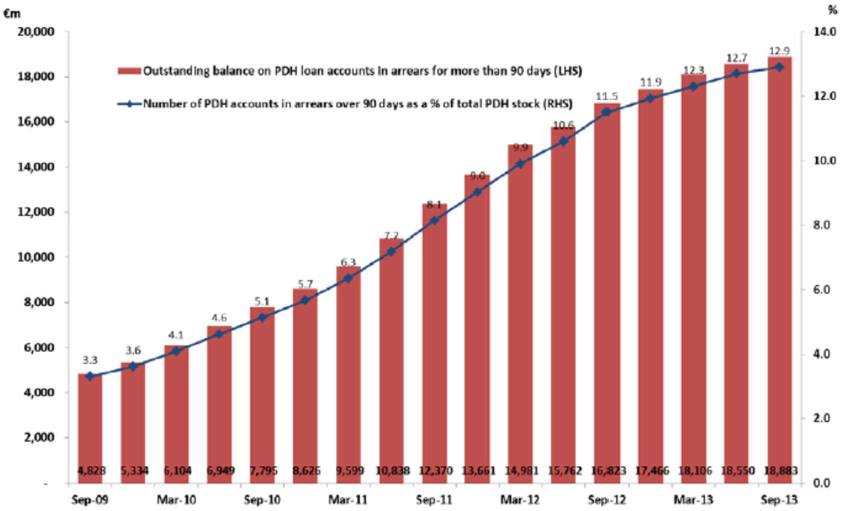


Figure 2.7 PDH Mortgage Accounts in Arrears over 90 days (end of quarter 3 2013)
Source: (Central Bank of Ireland, 2013)

Increases in mortgage arrears can be attributed to many factors. There has been widespread research covering the area with many focusing on high LTV values (Ellis, 2008b) & (Connor and Flavin, 2013); Reduction in disposable income (Kelly, 2012); Households with children (McCarthy, 2014) & (Keese, 2009); Unemployment (McCarthy, 2014), (Keese, 2009), (Kelly, 2012) & (Gyourko and Tracy, 2014); Mortgage repayments based on two incomes to sustain them, and the incomes staying at that level (McCarthy, 2014); New bankruptcy laws in the US (Li et al., 2010); Strategic defaults (Bhutta et al., 2010) & (White, 2010); Negative equity (Ellis, 2012), (Foote et al., 2008b), (Connor and Flavin, 2013), (Connor et al., 2012) & (McCarthy, 2014).

The drop in prices of Irish residential properties after the property bubble burst were extensive, with prices in Dublin and surrounding areas falling 55% and prices elsewhere in the country falling 43% (Kelly, 2012). This drop is second only to the severe drop in prices recorded in Japan 1991. Irish house prices reached their peak in late 2006, but many borrowers bought homes at inflated prices up until the end of 2007. In 2008 the prices really began to fall and only started to recover gradually from 2012 onwards.

The price increases in Irish property values were driven by a number of factors, such as the availability of credit and the lack of regulation by the regulator, and competition in the lending market domestically which unnaturally inflated prices. In the year 2000 there were twelve banks registered with the Central Bank of Ireland (CBI) who were allowed to lend credit in the form of mortgages. By 2008 this figure had increased to seventeen with a number of new and foreign banks joining the market (Norris et al., 2010). Anglo Irish Bank were an established Irish domestic bank, who had not previously engaged in mortgage lending; Bank Of Scotland Ireland and Danske Bank both joined the fray as well as both AIB and Bank Of Ireland setting up their own mortgage banks.

This increased competition in the market generated a race for market share between the banks, who reduced the margins on certain products in order to make the rates more appealing to borrowers (Norris et al., 2010). Some of the changes implemented by the banks during the early 2000s were interest only mortgages, tracker mortgages and higher LTV values and longer mortgage terms. Many of these changes have directly helped to increase the occurrence of mortgage arrears and payment difficulties in the Irish market.

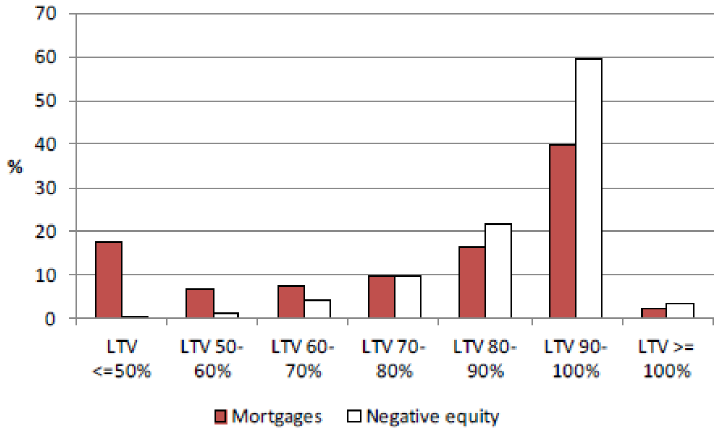


Figure 2.8 Distribution of mortgages & negative equity by initial loan-to-value ratio
Source: (Duffy et al., 2013)

Figure 1.4 details the percentage of mortgages in arrears by year of drawdown, and from this it is obvious that mortgages drawn down in 2006, 2007 & 2008 at the height of the property boom are the mortgages worst affected by mortgage arrears. Figure 2.8 displays the distribution of mortgages by LTV values for mortgages. Looking at each of the high LTV bands, the vast majority of the mortgages with arrears are focused here. Almost 60% of the loans with arrears are mortgage with an LTV value of between 90 and 100%.

Figure 2.9 displays mortgages split out by initial mortgage term and arrears. Roughly 50% of the mortgages that are in arrears have a mortgage term of between thirty one and thirty five years, while the more traditional mortgage term of twenty years or less has only approximately 5% of the total mortgages in arrears.

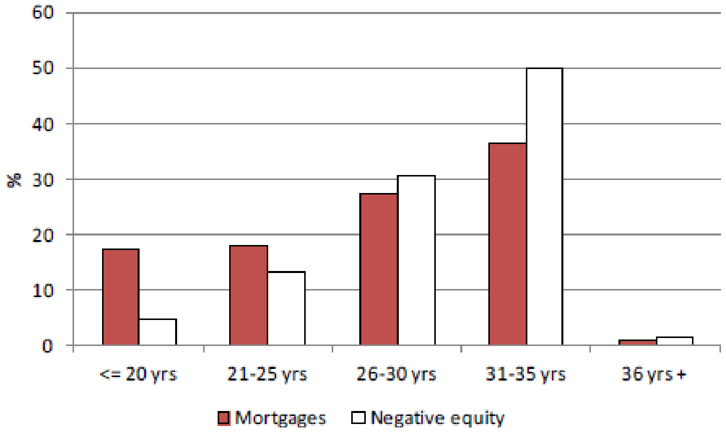


Figure 2.9 Distribution of mortgage & negative equity by original mortgage term
Source: (Duffy et al., 2013)

Banks who were seeking market share by offering lower mortgage rates and higher levels of funding at longer terms, have pushed borrowers ever closer to arrears. Before the housing boom there were little or no home loans with terms greater than twenty years, or with LTV rates of over 80%. Ultimately, now that there are a large number of mortgages in difficulty it is these mortgages that are making up the bulk of the arrears.

2.3.3 Causes of Mortgage Arrears

(Kelly, 2012) found that disposable income has dropped 8.1% from its highest levels in 2007. With lower levels of income households are less likely to be insulated against any income shock or unforeseen payment. When this happens it is likely that they may not be able to meet all of their credit obligations. If a household with a mortgage has sufficient disposable income to be able to save on regular basis, then they have a 3% lower possibility of defaulting than a household that is not able to maintain regular savings (McCarthy, 2014). However, there are many households that are not able to keep up with their payments and make savings, due to a number of factors.

During the height of the credit boom many of the mortgages that were taken out were reliant on two salaries to support the high payments, given the high property prices (McCarthy, 2014). As a result of the higher LTV values being allowed at the time homeowners were faced with large monthly payments to service a mortgage that were based on the two salaries staying at the same amount, or increasing. As employment conditions in the country declined rapidly from the onset of the financial crisis in 2008, income and employment terms decreased for many people. This put additional strain on households, and particularly young households who were burdened with long term mortgages with high LTV values. Even if all parties to a mortgages are still in full employment their salary may be lower, but the expectation at the time of taking out the mortgage would have been that their salary would only move in an upwards direction.

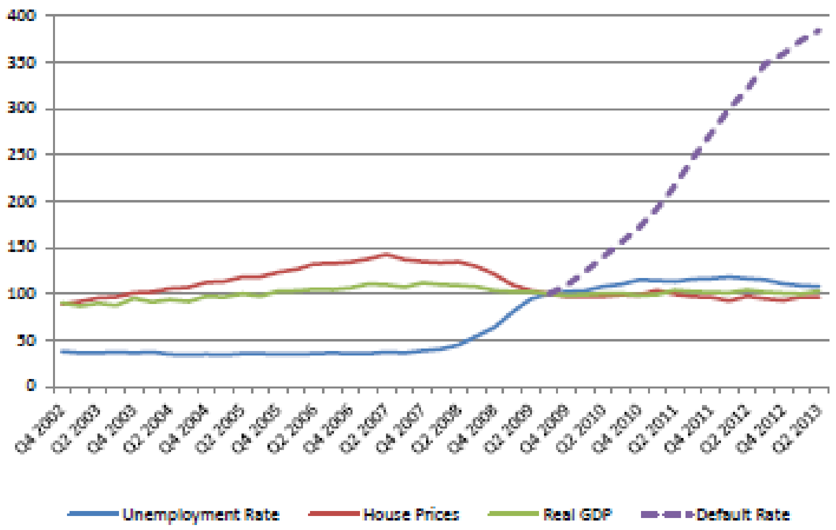


Figure 2.10 Comparative Time Series of Irish Residential Property Price Index, Unemployment Rate, Per-Capita Real Income and Home Loan Default Rate
Source: (Connor and Flavin, 2013)

Unemployment at a broad level has become a factor in the level of mortgage arrears. Figure 2.10 tracks the Irish unemployment rate, house prices, and per capita real income and mortgage arrears. It would appear that as the unemployment rate started to rise and as house prices and the per capita income started to fall, the rate of arrears in the country started to increase. It is hard to surmise when exactly the rate of arrears started to increase, as data is not available at a national level from prior to quarter three 2009.

National or even some local unemployment rates are not useful as an indicator of arrears when looking at an individual mortgage, but taken at a broader level the rise of unemployment from approximately 4.5% to 12% has been closely followed by a steep increase in mortgage arrears (McCarthy and McQuinn, 2011).

When looking at employment and individual mortgages, (McCarthy, 2014) found that national or regional levels of unemployment were not a good indicator of arrears, but if the head of the household is in a state of fragile employment, or has just been laid off then this is very much an indicator for mortgage arrears. In Ireland roughly 25% of households up to date with their mortgage payments have a head of household in a position of fragile employment. For households in arrears, this figure is closer to 40% (McCarthy, 2014). If the head of the household was to lose their job, then this would have a direct and adverse effect on the indebtedness of the household (Keese, 2009).

Figure 2.11 shows the close relationship between the unemployment rate and the rates of arrears in both the UK and Spain. In both countries, when the rate of unemployment dropped, the rate of arrears also dropped. Again, in both countries when the rate of unemployment increased, the rate of arrears also increased. While this does not necessarily show a causal factor, it would appear that the two are correlated.

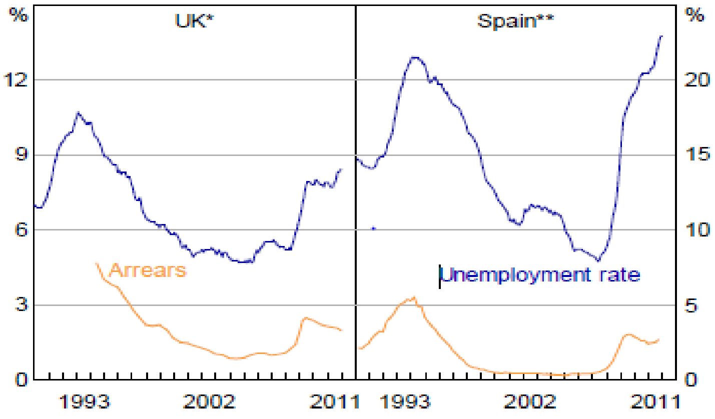


Figure 2.11 Mortgage arrears and unemployment in the UK and Spain
Source: (Ellis, 2012)

(Kelly, 2012) infers that unemployment levels have a much greater effect on arrears in Ireland than the drop in house prices. He states that the ability of a homeowner to service their debts is one of the most important factors in working out arrears. In order for most homeowners to be able to pay their credit debts, they need to be in employment, so he believes that policies that look to get the domestic economy back in order, and hence increase employment rates, are much more likely to result in a decrease in arrears than policies that focus on debt reduction. (Gyourko and Tracy, 2014) find that unemployment most certainly should play a role in prediction models, but it should be borrower level employment/unemployment detail and not proxy rates. They find that proxy rates of unemployment do not improve the accuracy of models of default.

A lot of the current research on mortgage arrears and default focuses on the role of negative equity in causing arrears. (Gerardi et al., 2008) infer that a household that has experienced a 20% decrease in the price of their property is fourteen times more likely to default on the loan than a household that has experienced a 20% increase in the price of the property. A homeowner who has suffered a loss of 20% on the value of their home is unlikely to be able to sell the property to discharge their mortgage if they are in difficulty with the payments, whereas a homeowner in positive equity could sell the property to discharge the loan, or at least release some of the equity in the home to help. (McCarthy, 2014) found that the proportion of borrowers in distress on their mortgage who have positive equity in their home is 57%, whereas the proportion of homeowners who are current with their mortgage and have positive equity is 64%.

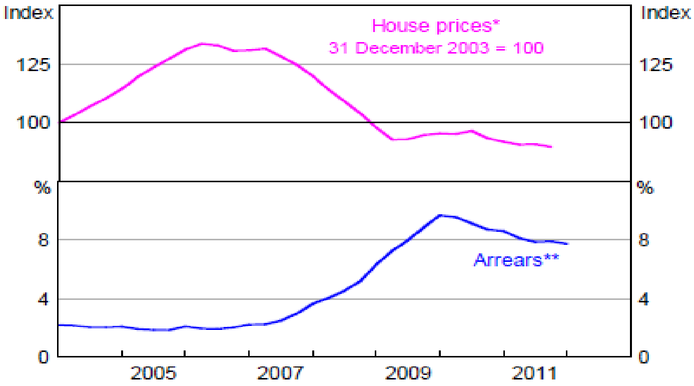


Figure 2.12 US house prices and mortgage arrears
Source: (Ellis, 2012)

House prices are not directly linked with the level of arrears, though it would seem that there is a correlation between the two. Figure 2.12 shows the rate of house price growth in the US during the period 2004 to 2011 as well as the arrears rate. The rate of arrears stays constant all during the time that house prices were increasing and only starts to increase once house prices start to decline. There may not be a direct causal factor between the two, but they certainly seem to be interlinked.

The initial LTV of a mortgage has a direct effect on the balance of the loan and the amount the borrower has to be repay, and mortgages with high starting LTV rates have a higher susceptibility to negative equity and default (Ellis, 2008b) & (Connor and Flavin, 2013). Negative equity is present in a lot of borrowers who go in to arrears/foreclose, but a mortgage being in negative equity is not necessarily going to be in arrears. Most mortgages taken out in the recent past in Ireland will be in negative equity, and given higher prices, longer terms and higher LTV values at origination then negative equity is widespread. One estimate of negative equity in Ireland is 50% (McCarthy, 2014), but the highest rate of arrears so far in the country has peaked at 18.54% which is a long way off 50%. There are also many borrowers who are in positive equity in their homes who are facing arrears, which would show that negative equity is not the overriding reason for arrears, though it may be one of the factors involved.

There are many life stages and events that can put financial pressure on a household such as a wedding or having children. Both (McCarthy, 2014) & (Keese, 2009) find that households with children are much more likely to be subject to over indebtedness and arrears. (Keese, 2009) finds that childbirth greatly increases the chance of a household getting in to a situation of over indebtedness, while it also raises the probability of a decline in the performance of the household’s debt. (McCarthy, 2014) states that households with dependent children are 5% more likely to be in arrears that households with no dependent children.

Some borrowers might find strategic default an attractive option when faced with paying substantially higher mortgage payments than they would for a similar house if they were renting it, even when they are able to afford the payments for the mortgage. (Bhutta et al., 2010) found that in Palmdale California, the cost of servicing a mortgage on a 3-4 bedroom house that would have fallen in value from $375,000 to $200,000 in the property crash, would be about $2,500 a month. Similar houses are demanding a premium of approximately $1,300 a month in the rental market. They surmise that unless the homeowner believes that house prices are going to recover very strongly, that there is no reason for the homeowner to continue paying their mortgage, when they are paying such an obvious premium to do so. (White, 2010) states that there are millions of homeowners in the US who are underwater on their mortgages, who could save thousands of dollars by strategically defaulting on their mortgage payments. White says that the average US household who purchased a home at the peak of the property bubble and is in a large amount of negative equity, cannot expect to get a return on their investment for a very long time and maybe not even in their lifetime.

(White, 2010) states *"In short, the financial costs of foreclosure, while not insignificant, are minimal compared to the financial benefit of strategic default, particularly for seriously underwater homeowners."*

The only surprising element to defaults and strategic defaults for White, is that there aren't more US homeowners who are strategically defaulting. There are many reasons he gives for homeowners not strategically defaulting on their loans, some of which are social stigma; the inability to be able to carry out complex calculations to show the cost of defaulting against the cost of continuing to pay the mortgage; and the fact that most homeowners want to believe that their house is worth much more than it actually is.

While White believes many homeowners will not default due to the social stigma attached, (Guiso et al., 2013) found that homeowners are much more likely to default if they knew of someone else who had already defaulted. In part this can be attributed to the fact that the social stigma attached is lessened, but another possibly more important reason is that the homeowner will likely now know the true cost or savings involved in defaulting. Policy makers are concerned that this may be the case, whereby once a number of people in a neighbourhood have defaulted, all other homeowners may follow in what would be seen as social contagion (White, 2010). This has shown to be the case by (Guiso et al., 2013) whereby clusters of homeowners who have defaulted exist in certain neighbourhoods, and this is certainly a worrying trend for lenders.

For borrowers with low credit ratings, the cost of default is not as high, as their credit rating would have signified that their repayment propensity was already low. However, for borrowers with good credit ratings, the cost of default is much higher, as they will lose their good credit rating, and face higher future borrowing costs as a result (Bhutta et al., 2010). They estimate that in the sample of mortgages they used, there was a strategic default rate of 20%, but there were large differences across the different types of borrowers who strategically defaulted. They found that the median borrower defaults at -62% negative equity, though borrowers with lower credit ratings defaulted at -51%, while borrowers with a good credit rating didn’t default until -68% negative equity.

(Bhutta et al., 2010) stipulate that the gap between where borrowers with bad credit ratings will default compared to borrowers with good credit ratings can be explained by the future cost of funding that both groups will face after defaulting on a mortgage. For the borrowers with bad credit ratings they may only face a slight increase in funding costs, whereas the borrowers with good credit ratings will face much higher costs of funding after defaulting. In the US it is still possible for an individual who has a bad credit rating to get a mortgage, but if you have just defaulted on a loan then you will likely have to wait a number of years before you apply for another. (White, 2010) believes that there are many borrowers in the US who have decided to default on their home loans as a result of the premium they are paying over market values to stay in their homes. Some of these strategic defaulters have even gone out and gotten a second mortgage on another property before defaulting on their first mortgage, so that the decline in their credit rating won’t affect their ability to get another mortgage.

2.4 Addressing Arrears

Recourse is a powerful tool that ultimately stops more borrowers from defaulting on their loans. If a lending institution has recourse on a loan, then they have the ability to recoup all of their losses from a borrower in the event of a default, against both the security provided on the loan and against the borrower themselves. In the case of non-recourse loans the lender can only recoup their costs against the security provided on the loan. In the case of mortgages this generally means just the property the loan is secured against. In the US it is commonplace that mortgages may be non-recourse, where the lender may only recoup their losses by repossessing the property and selling it off.

(Ghent and Kudlyak, 2011) found that in general, the probability that a borrower will default is roughly 20% higher in states where the lender has no recourse, than it is in states where the lender has full recourse over the borrower. The rate of default probability gets higher for mortgages of higher value, with borrowers who have mortgages between $750,000 and $1 million, 66% more likely to default. This shows that recourse can be a very powerful tool for lenders in helping to keep down mortgage default rates, as borrowers are much less likely to default on their loans. In Ireland and most European countries, lenders have full recourse on any mortgages they provide. This means that they will try to recover any losses not covered by the sale of a repossessed property. As a result borrowers in Ireland would be much more reluctant to default on their mortgage, especially when they know the value of the property falls well short of the outstanding value of the loan it is secured against.

Mortgage foreclosures or repossessions are common in the US market, where a mortgage lender will repossess the house in order to be able to recoup as much as the costs as possible. In Ireland though the level of house repossessions is much lower. Traditionally banks have been slow to repossess a property unless it is the absolute last resort. Figure 2.13 shows the cumulative number of residential properties in possession by the banks in Ireland, and the number of properties being added each month by banks is very low compared to other countries. As part of the government guarantee of September 2008 the banks had to agree to a 12 month moratorium on all repossessions, though it is always a last resort for them with very high legal and transaction costs associated with repossessing a house.

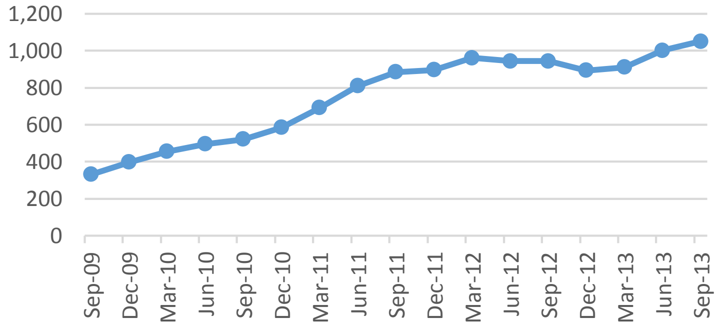


Figure 2.13 Residential properties in repossession quarter 3 2009 - quarter 3 2013
Source: (Central Bank of Ireland, 2013)

The preferred method of choice for both Irish mortgage lenders and the Irish government, is that mortgages that are in difficulty for a prolonged period of time are restructured. Re-structuring may take many forms including an interest only term; a payment moratorium; a change in the amortising schedule so that the borrower only pays interest for a certain period of time or a write-down of some of the outstanding balance

(Central Bank of Ireland, 2013). At the end of quarter 3 2013 there were a total of 80,555 mortgages in Ireland that had been re-structured in some form. This represents almost 57% of all loans in arrears, and approximately 10.5% of all outstanding mortgages.

2.5 Conclusions

This chapter has focused on some of the current literature available on mortgage arrears and some of the many factors that are affecting the rate of arrears and defaults. In order to be able to understand what is happening with the current spate of arrears, defaults and repossessions it is important to look back and see what has contributed to and shaped the mortgage market in the past ten to fifteen years.

In both the US and Ireland, and many other countries there has been a property bubble that has contributed to house prices growing at near exponential rates. As a result the debt burden has greatly increased for homeowners, and as a result of the financial crisis many of these homeowners are experiencing difficulty paying their debts. Many homeowners who bought at the height of the housing markets are young people - typically the most active users in the housing market. Negative equity will not allow these young homeowners to move from the properties they are currently in. Lack of movement in the housing market has perpetuated a stagnation in prices, and ultimately prices fell substantially from their peak which further increased negative equity.

Job losses have also had a profound effect on the ability of many households to pay their mortgage, as have wage cuts and uncertainty in the jobs market. Many mortgages were given out on the premise that two salaries would be servicing the debt, but in some cases one or both salaries have been decreased or cut altogether and this has greatly increased the stress and strain on households with large amounts of debt.

The current research on mortgages in arrears/default shows that the problem of homeowners in difficulty is a multi-faceted one, with many reasons for a borrower to be in distress. A sample of the reasons why a homeowner might be in difficulty is as follows: increased negative equity on the property; unemployment causing loss of income; higher LTV values and longer mortgage terms; Life events such as the birth of a child or other income shocks such as health costs which can be hard to quantify.

3. LITERATURE REVIEW PREDICTIVE MODELLING

3.1 Introduction

This chapter will cover related research literature in the field of predictive modelling and data mining. Algorithms used in predictive models will be discussed in brief detail, along with examples of how each can be used. The key methodologies for building predictive models will be covered, along with examples of models that have been built to predict arrears and mortgage default. Different techniques for evaluating these models will be discussed, to help decide how the experiment should be evaluated.

For many predictive models one of the outcomes is often rare, such as is the case in most mortgage arrears models. In the case of models built for predicting mortgage arrears there tends to be much higher degree of mortgages that are not in arrears, and have not defaulted. As a result these datasets have a much higher concentration of good loans than bad loans. This can prove difficult to model, as the bad loans are very much outnumbered. Current research on addressing this imbalance within the model will be reviewed, including over-sampling, under-sampling and other relevant methods with the intention of implementing some of the methods in to the model being built.

3.2 General Predictive Models

> *"Knowledge discovery is the nontrivial extraction of implicit, previously unknown, and potentially useful information from data."*
>
> (Frawley et al., 1992)

(Fayyad et al., 1996) defined a process that they called Knowledge Discovery in Databases of KDD for short, which is a framework for extracting useful information from databases. One of the steps of the KDD process is Data Mining, of which predictive modelling is but one of many types of data mining. Predictive models seek to be able to predict the future behaviour in a certain situation, given what has happened in the past. Most predictive models are trained using historic data, but then are tested against real world data to ascertain how well they perform.

There are many applications for predictive modelling including medical applications such as predicting cancer rates (Delen et al., 2005), predicting the spread of oil slicks in the case of spills or leaks (Liu et al., 2011), predicting the outcome of elections (Tumasjan et al., 2010) and more recently predictive modelling has become prevalent in many sports (Silverman and Suchard, 2013) & (Stekler and Klein, 2012).

Predictive models, when applied correctly can greatly enhance the knowledge about a certain subject and can reduce risk, increase profits and offer greater products or services. Models have been used in banks for many years to model the inherent risks associated with lending, so that the banks can work out which customers are likely to be able to repay debts, and which are not.

In recent times with the tightening of credit in the Irish market, the credit risk models have become more and more important, so as to reduce the risk the banks are taking on. However there has been less focus on modelling the current balance sheets of the banks, and how they are performing. With the recent escalation in number of people going into arrears, predictive modelling on the loan books and specifically the mortgage loan book, would help to reduce losses being incurred if some of these losses could be prevented.

3.3 Imbalanced Datasets

For many classification and prediction models, imbalanced datasets can cause an issue and reduce the effectiveness and accuracy of the model. In many cases one of the outcomes or classes is much rarer than the other, and this gives rise to a number of issues. If the average rate of default or non-payment in loans is 3%, and a bank has a loan book that contains 10,000 loans, then 300 of the loans in the loan book will default. However, 9,700 of the loans in the loan book will not default. If the bank would like to build a predictive model to work out what loans are going to default then they need to do something to address the imbalance of class in the dataset.

Simply put, if the model used by the bank said that every loan was a performing loan, and that none of the loans were going to default, then the model would be correct 97% of the time. On the face of it, 97% is a high accuracy level. However, it would be incorrect 3% of the time, and in all of these cases it would have predicted that a loan that actually defaulted was not going to default. This can have many implications for an institution such as a bank – not least of which is having the correct capital requirement in place for non-performing loans. In most prediction models there is generally a cost associated with misclassifying a record with the wrong class, though in the case of many models such as a loan default model the cost associated with the different types of misclassification can be quite substantial (He and Garcia, 2009). For the bank default prediction model mentioned earlier, the cost of misclassifying a good loan as a bad loan would carry little cost, though misclassifying a bad loan as a good loan would have a substantial cost associated. When this cost is taken in to account the overall accuracy of the model would in effect be much lower than the earlier reported 97%.

(Chawla et al., 2011) describe a dataset as being imbalanced if the number of records associated with each class is not approximately equal within the dataset. This would dictate that the ratio between classes would have to be close to 50/50 for a dataset with two classes. In reality though one outcome often occurs considerably less often than the other. Most classification or prediction algorithms are designed so that they will work well with a dataset that is balanced (He and Garcia, 2009), but when the dataset is skewed towards one of the classes, then the model performance may decrease significantly. For imbalanced datasets the algorithm used tends to focus more on the majority class at the expense of the minority class. To address this scenario there are a number of measures that can be put in place to prevent this. Some of the techniques used to redress imbalanced datasets will be discussed in the following sections.

3.3.1 Random undersampling and oversampling

Random undersampling and oversampling seek to address the imbalance in a dataset by either reducing the number of items in a dataset in the majority class, or increasing the number of items in the minority class. Random undersampling takes a random sample of records from the original majority class within the dataset and uses these records to build the model instead of the all items from the majority class. This has the effect of balancing the ratio of minority to majority class items.

Oversampling works on the minority class and adds additional minority class items in to the dataset to even out the balance in the classes. Random oversampling takes a random sample of the minority class and adds these records to the minority class along with the existing records. This has the effect of duplicating some/all of the minority class records in the dataset, depending on the level of oversampling. Oversampling increases the overall number of records in the dataset, while undersampling decreases the total number of records in the dataset (He and Garcia, 2009).

Both random oversampling and undersampling can help to equalise the imbalance in the dataset, but both are not without their issues. By undersampling the majority class it is very possible that important concepts for the majority class may be removed from the model, and by utilising random oversampling the chance of overfitting the minority class is a distinct possibility (Chairi et al., 2012).

3.3.2 Informed Undersampling

Informed undersampling seeks to use sampling to decrease the number of majority class items in a dataset, though it does it in a very different way to random undersampling. There are a number of informed undersampling methods available, but two of the most popular are EasyEnsemble and BalanceCascade (He and Garcia, 2009).

EasyEnsemble is an unsupervised algorithm that samples the majority class of a dataset and creates an number of samples which are then all combined together to create an ensemble, which is boosted using the AdaBoost algorithm to create a classifier (Liu et al., 2009). The effect of using the boosting algorithm and a bagging-like activity has the desired effect of creating a classifier than generalises well, and produces a good sample of the majority class. BalanceCascade is different to EasyEnsemble as it takes a supervised approach to the sampling of the majority class, but works in a very similar way to work out which records from the majority sample should be included.

Another variation of informed undersampling is a K-nearest neighbour (KNN) classifier to identify the records to be used as part of the sample. There are at least four variations of KNNs that can be used for informed undersampling, and all of them use a variant on the distance from the majority class items to the minority class items to work out what items to select for the sample. According to (He and Garcia, 2009) the NearMMiss-2 method show good promise at providing good results for imbalanced learning.

3.3.3 Synthetic Sampling

Synthetic sampling methods look to resolve the lack of minority samples in the dataset by adding new samples in to the dataset for the minority class. Unlike random oversampling these new data items that are added to the dataset are synthetic samples that have been created to be similar, but not identical to the pre-existing minority data items. Synthetic Minority Oversampling Technique (SMOTE) is an example of a synthetic sampling method that creates synthetic minority class data to add to the dataset.

SMOTE was proposed by (Chawla et al., 2011) who created synthetic data by generating data that sits somewhere between a data item and it's K nearest neighbours. In the implementation described, they require 200% oversampling which means that the number of minority samples in the dataset needs to be tripled. This means that for each data items in the original minority class there are two new synthetic data items created, where the synthetic data items sit somewhere between the original data item and two of its nearest neighbours (Chawla et al., 2011). The effect of SMOTE is that the model being trained on the dataset should generalise much better for the minority class, as the synthetic data is similar to the existing minority class data, but is not identical. Figure 3.1 shows how a synthetic example is created for the original data item x_i. The six nearest neighbours for x_i are identified in Figure 3.1 (a) and then in Figure 3.1 (b) the new synthetic data item is created between x_i and x^$_i$. If more than one synthetic item was required the new item would be created between x_i and another of its nearest neighbours. The main drawback to using SMOTE is that the model generated may over-generalise due to the synthetic data that has been added in.

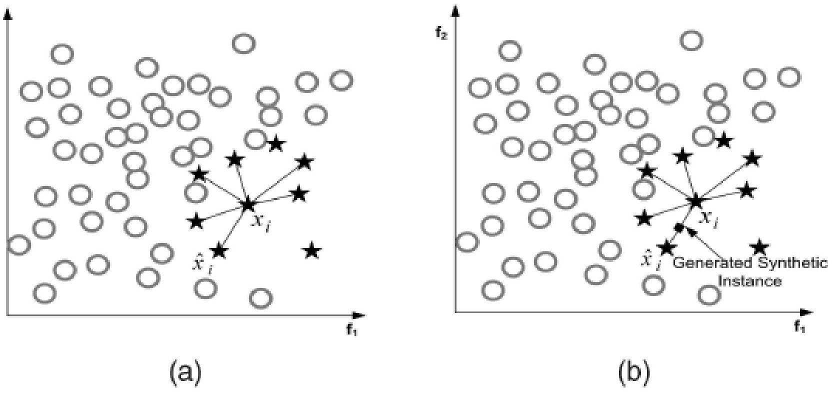


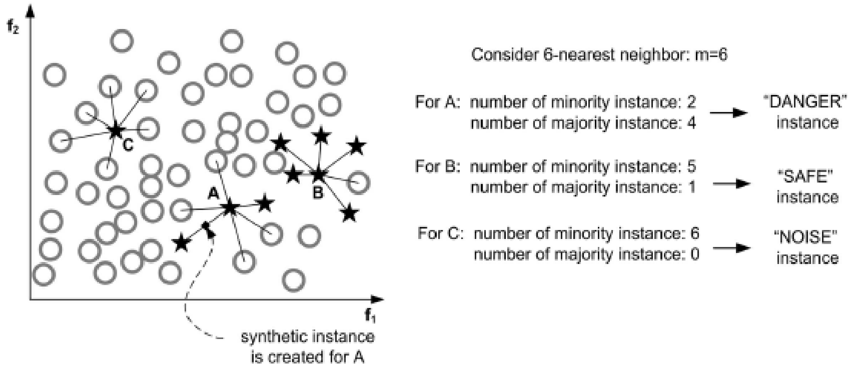
Figure 3.1 Creation of Synthetic data using SMOTE
Source: (He and Garcia, 2009)

3.3.4 Adaptive Synthetic Sampling

Given that there are certain limitations to SMOTE, there have been a number of adaptive synthetic sampling methods put forward in recent years. The main reason that SMOTE can over-generalise is that it creates the new synthetic data items regardless of what class the nearest neighbours belong to, which can lead to issues of overlapping between classes (He and Garcia, 2009).

Borderline-SMOTE is one of the Adaptive Synthetic Sampling methods introduced to rectify some of the issues with the original SMOTE methodology for oversampling. It looks to focus on the data items in the minority class that are closer to the majority class, as opposed to all of the items in the minority class (Han et al., 2005). The reasoning for this approach is that the data items in the minority class that are most difficult to classify, are the items that are closest to the majority class. By creating synthetic data only for these 'borderline' data points, then the algorithm use for the classification or prediction should be better able to generalize and predict these values (Han et al., 2005).

The Borderline-SMOTE methodology identifies the 'borderline' data items that are in the minority dataset which have both minority class neighbours and majority class neighbours. If a data item in the minority class has only majority class or only minority class neighbours, then this data point will not be used in the creation of new synthetic data (Han et al., 2005). Only the data points that have neighbours in both classes where the number of each is comparable will be included. In Figure 3.2 the data point A is said to be in the 'DANGER' set, which is the dataset where the items have both classes as neighbours. This is one of the items that will be used to create synthetic data, whereas points B and C will not (He and Garcia, 2009).



f₂
C
A
B
f₁
synthetic instance is created for A
Consider 6-nearest neighbor: m=6
For A: number of minority instance: 2
number of majority instance: 4
"DANGER" instance
For B: number of minority instance: 5
number of majority instance: 1
"SAFE" instance
For C: number of minority instance: 6
number of majority instance: 0
"NOISE" instance

Figure 3.2 Generating synthetic samples using Borderline-SMOTE
Source: (He and Garcia, 2009)

Figure 3.3(a) shows an example of a dataset with an imbalance towards the majority class. Figure 3.3(b) shows the 'borderline' points that have been selected by the Borderline-SMOTE algorithm to use for producing the synthetic data items, and Figure 3.3(c) shows the dataset with the new synthetic items added in. These new synthetic items are primarily focused around the border of the two classes, so this should help with the generalisation of the model for the 'borderline' items in the minority class.

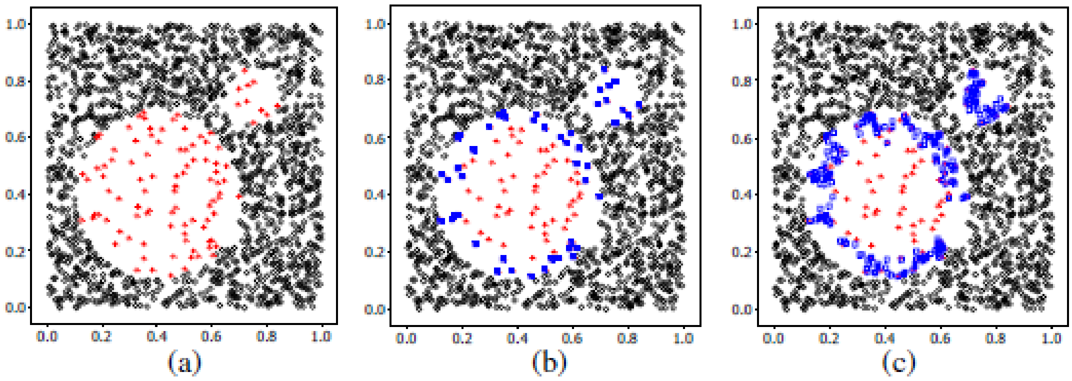


Figure 3.3 Borderline-SMOTE example
Source: (Han et al., 2005)

ADASYN seeks to build on the need to generate more relevant sampling data than just generating synthetic samples for all data items in the minority class. In many ways ADASYN is quite similar to Borderline-SMOTE as it creates synthetic data based on the items in the dataset that are likely to be the hardest to classify. The method uses a weighted distribution to work out which data items need to have more synthetic items created, based on the distribution of the minority class (He et al., 2008). For data items that would prove more difficult to classify there are more synthetic example created, and for those that are easier to classify there are less synthetic examples created. ADASYN as a result addresses the class imbalance in the whole dataset, but also compensates for skewed distributions within the minority class itself (He and Garcia, 2009).

3.3.5 Cost Sensitive Methods

Another way to focus on imbalances in datasets is to use a Cost Sensitive approach to modelling the data. Cost sensitive methods concentrate on applying a cost to the misclassification of data, rather than creating additional or synthetic data items. The use of a confusion matrix is central to how most cost sensitive methods operate, and takes in to account how much each misclassification is likely to cost. There is likely to be little cost associated with classifying a majority class item as a minority class item, though if a minority class example is classified as a majority class item there is generally a much higher cost associated (He and Garcia, 2009). Generally there is no cost associated with a correct classification, so items correctly classified no not affect the overall costs.

Cost sensitive methods can be applied in a number of ways. Three examples of how they can be applied are as follows: the misclassification costs are used to select the best training distribution in order to minimise the associated costs; Adaptive Boosting is used to assign a higher cost to the misclassified examples in a dataset so that these items are continually the focus of the training algorithm; cost sensitive methods can be applied to decision trees to work out what the decision threshold should be or how to split the data at each node (He and Garcia, 2009).

3.4 Mortgage Predictive Models

There has been huge focus on modelling mortgages in recent years with the advent of the financial crisis and the decline in house prices in many developed countries, coupled with the steep rise in mortgage arrears and default rates. Much of this focus has been on the US housing market as this was seen by many as the precipitating factor for the financial crisis that has emanated around the world since 2007. Much of the research available focuses on trying to predict what mortgages are likely to go in to arrears or default, with some of the literature trying to work out if the mortgage bubble could potentially have been predicted and hence, somewhat avoided.

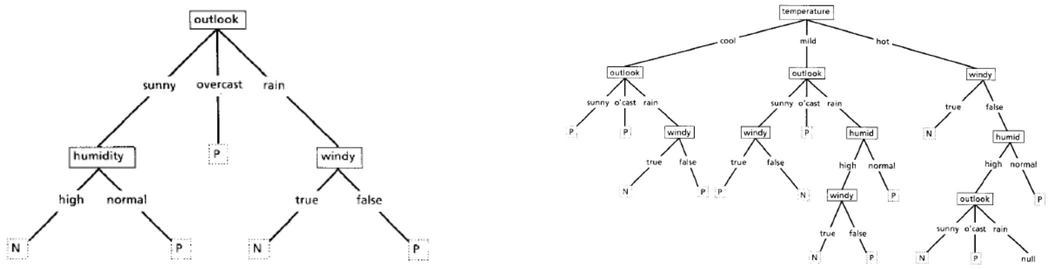
In many ways predictive models are very similar to classification models, as the model seeks to predict what class each record belongs to, albeit what the future class is likely to be in the case of most predictive models. In this case both classification and predictive models work in the same way, but classification models can take many forms that are different to predictive models such as clustering.

Researchers have taken many different approaches to modelling arrears and default, with most of the common algorithms used to predict what mortgages are going to go bad. Using Neural Networks is one of the most common approaches taken with examples of this being (Scheurmann and Matthews, 2005), (Odeh et al., 2011), (Jagannatha Reddy and Kavitha, 2010) and (Hassan and Abraham, 2013). Other approaches have proven popular with Support Vector Machines (SVMs) accounting for a large part of the research (Wang et al., 2007), (Shin et al., 2005) & more.

3.4.1 Decision Trees

Decision Tree algorithms seek to model data based on the inherent characteristics of the inputs in the data. The algorithm tries to break the data down in to progressively smaller sets of data, by splitting the data based on the characteristics of the input. When the decision tree splits the data in to smaller groups, it is trying to separate the classes, so that ultimately each smaller group of records should belong to one class. In the case of most predictive models this will mean that the data should be broken down in to a positive and negative class, which will ultimately be assigned to each of the smaller groups or leaves as they are commonly known.

With each step of the decision tree, the algorithm looks to find the best splitting point for the data by selecting the best attribute to split the data, so that the predicted classes are kept separate as much as possible. In order for this to be achieved, the attributes that form the dataset must provide some information that will allow the data to be separated into the distinct classes. If there are two records in the training data being used that have the same attributes, but are of differing classes, then a decision tree will not be able to accurately split the data, as it will not know which class to assign to the records (Quinlan, 1986). In the case of this happening, then the class assigned will be the class that holds majority in this leaf (Friedman et al., 1996). If the number in each class within the leaf is even, then the majority class of the parent node from which the leaf was formed is chosen.



(a) (b)

Figure 3.4 Simple decision tree structure (a) and complex decision tree (b)
Source: (Quinlan, 1986)

Decision trees can be very simple and generalize well for training and testing data sets, or they can be very specific for a training set and not perform quite so well for a testing dataset. Occam's razor states that in the case of two decision trees (or other algorithm) with the same generalization error and/or classification error, the simpler of the two trees should be chosen as the tree to be implemented (Domingos, 1999) & (Blumer et al., 1987). The basis for this is that the simpler tree will be much better at generalising for unseen data, and the computational cost should also be lower. In the case of Figure 3.4 (a) and (b), if both the decision trees had comparable misclassification rates and generalizing rates, then the tree in Figure 3.4(a) should be chosen as the preferred model as it is simpler and will also likely generalise better.

There are a number of issues with decision trees as per (Friedman et al., 1996). One of the drawbacks is that decision tree algorithms generally apply a top down approach, and when splitting the data at each of the nodes the algorithm only looks ahead one step. As a result of only being able to look ahead one step, the algorithm may not choose the optimal split for a node as each attribute is taken in to account separately. In most cases in a dataset, a single attribute may have no bearing on the target value, but a number of attributes in combination may be a strong predictor. Decision trees are also not very adept at handling missing values, or values that were not seen in the training dataset.

Decision Trees are not commonly used in arrears prediction models for some of the reasons mentioned above. However, in most cases they are easily explainable, and the list of rules used for splitting the data at each of the nodes can easily be re-produced so they certainly have their merits.

3.4.2 Neural Networks

> *"The brain is a highly complex, nonlinear, and parallel computer (information-processing system)."*
>
> (Haykin, 1994)

The advent of Neural Networks in the field of Artificial Intelligence and Data Mining came about when researchers wanted to see if they could model and train computers to work in a similar way to the human brain and its vast array of neurons. The human brain is vastly more powerful than any computer available, and can carry out much more complex tasks at a much greater speed. Computers are very efficient at processing large amounts of data in a short time, but ultimately they are very limited in the capabilities of learning and gaining experience from past activities.

With the introduction of Neural Networks though, computers are much better equipped to process data in a similar way to a human brain. Neural networks are different to other predictive models due to their parallel computing powers and ability to generalise on new data (Haykin, 1994). Another facet of Neural Networks is their ability to work out nonlinear functions that are inherent in a dataset that would not be easily discernible using more traditional algorithms (Kaastra and Boyd, 1995).

Figure 3.5 shows the makeup of a traditional Neural Network which has an input layer with one hidden layer and one output layer. The hidden layer is made up of neurons that represent the neurons of the human brain. Each of the nodes of the input layer joins to all of the neurons in the hidden layer, which in turn all join to the output layer. A connection point or "synapse" a weight is applied to each of the connections, which will be used in the computation of the output of the Neural Network. One of the redeeming features of Neural Networks is the ability to change the connection weights in order to recalculate the output, and hence learn from the data and generalize better (Haykin, 1994). In this way Neural Networks can perform in a comparative way to the human brain be learning from experience, though they are still not as powerful.

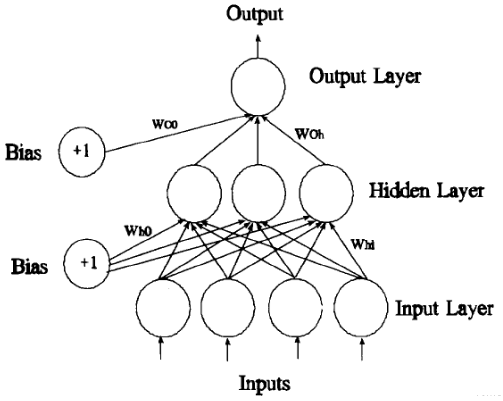


Figure 3.5 Neural Network example
Source: (Kaastra and Boyd, 1995)

Many researchers have made use of the power of Neural Networks when modelling mortgage credit. The versatility of Neural Networks means that they are able to learn from a dataset and should be almost universally applicable to most classification and predictive modelling tasks in some way.

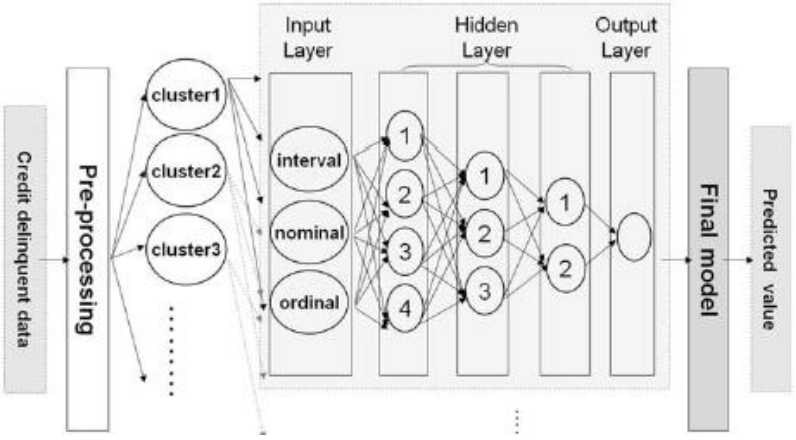


Figure 3.6 Locally Transductive Multi-layer Perceptron
Source: (Heo et al., 2009)

(Heo et al., 2009) proposed using a two-step model for predicting credit delinquents, by combining clustering with a multi-layer perceptron Neural Network. The clustering phase of the model divided the data in to ten distinct clusters which allowed the reduction of the number of input variables for the Neural Network which helped reduced the dimensionality of the data required for the prediction part of the model.

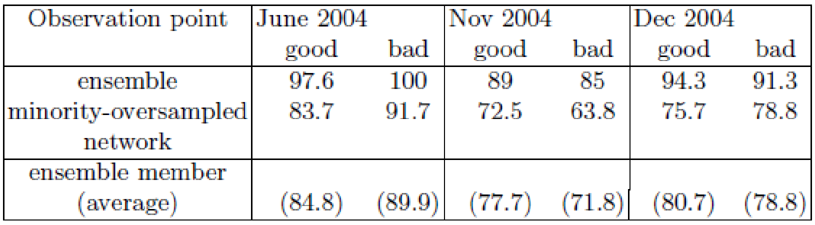
The Multi-layer Perceptron is a variation of a Neural Networks that is a feed forward network that uses back propagation to train the model by updating the synaptic weights in order to change the predicted values output. For the model created by (Heo et al., 2009) they have chosen a perceptron with three hidden layers, as they found this to be the optimal number of layers for generalisation, without overfitting the model. Their combination model of the clustering and Neural Network performed well compared to simply running the Neural Network by itself, and the time taken to achieve a good training score was much less for the combined model.

However, the overall prediction accuracy of the combined model was still relatively low. As with most financial datasets the dataset was imbalanced, with a much greater majority of non-delinquent loans in the dataset compared to delinquents. In the overall dataset there were only 684 delinquents which equates to approximately 3% of the overall dataset. The model performed well on predicting the non-delinquents with just over 98% of them classified correctly. For the delinquents though the accuracy rate was much lower at just under 52%. It would seem apparent that there should have been some form of sampling method used to redress the imbalance in the minority class in the dataset or to create synthetic data as discussed earlier in Section 3.3.

(Scheurmann and Matthews, 2005) modelled the loan data from a financial institution in Australia using an ensemble of Neural Networks, to test the common theory that an ensemble of classifiers will always work better than a single classifier. Their data, like many financial datasets was highly imbalanced towards accounts that were not in arrears, with just over 4% of all accounts in the dataset actually in arrears. Their findings were consistent with the literature, which states that in datasets that are largely imbalanced, the algorithm learns that majority class very well, as the expense of the minority class (Scheurmann and Matthews, 2005).

As part of the experiment they tried a number of different methods of increasing the training and classification of the minority class, including minority oversampling and majority undersampling. The minority oversampling method gave the best results of all the methods tested, and when viewed against a number of different single Neural Network classifiers trained on the original dataset, the minority oversampled dataset performed much better. However when the ensemble of classifiers was tested on the minority over-sampled dataset against the single classifier using the minority oversampled dataset, then the ensemble performed much better with previously unseen data as shown in Table 3.1. The ensemble network performed particularly well at correctly classifying the minority class, which in most classification or prediction models is crucial, as this will be the most important aspect of the model.

Table 3.1 Ensemble Neural Network vs Minority Oversampling performance
(Scheurmann and Matthews, 2005)



Observation point	June 2004		Nov 2004		Dec 2004	
	good	bad	good	bad	good	bad
ensemble	97.6	100	89	85	94.3	91.3
minority-oversampled network	83.7	91.7	72.5	63.8	75.7	78.8
ensemble member (average)	(84.8)	(89.9)	(77.7)	(71.8)	(80.7)	(78.8)

Many researchers have carried out work on using Neural Networks in comparison with other algorithms for feature selection in order to speed up the classification time for the neural network, and to reduce the issue of dimensionality. (Oreski et al., 2012) carried out an experiment using a Genetic algorithm for feature selection as part of an iterative model with a Neural Network for the classification. The process is shown in Figure 3.7 which exhibited promising results, and certainly decreased the complexity and computational cost of the model.

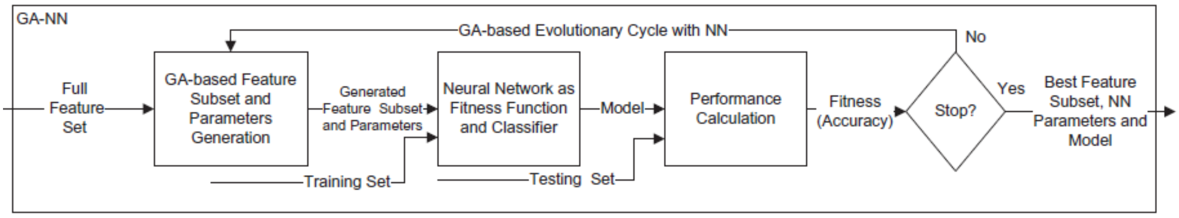

GA-NN
GA-based Evolutionary Cycle with NN
No
Full Feature Set
GA-based Feature Subset and Parameters Generation
Generated Feature Subset and Parameters
Neural Network as Fitness Function and Classifier
Model
Performance Calculation
Fitness (Accuracy)
Stop?
Yes
Best Feature Subset, NN Parameters and Model
Training Set
Testing Set

Figure 3.7 Genetic Algorithm with Neural Network
Source: (Oreski et al., 2012)

Neural Networks are sometimes viewed with a lot of uncertainty due to their "black box" modus operandi, whereby it can be very hard to tell what is actually happening in the algorithm, and how they are classifying data. For many users it can be very hard to accept that the model works on the data correctly, as they are not able to see the rules that were used for the calculations. (Setiono et al., 2008) proposed an algorithm called Re-RX which extracts a set of rules from a feed forward Neural Network, that are very similar to the rules created by Decision Trees. These rules mimic what is happening between the input layer and output layer of the Neural Network, so can be a useful tool for explain how the model reaches its conclusions.

3.4.3 Regression Models

Linear Regression and Regression models in their simplest form, are the modelling of (linear) relationships between variables. The most basic occurrence of a Linear Regression model where there is one independent variable and one dependent variable, the data is modelled using the function $Y = f(X)$ (Neter et al., 1996). For the example in Figure 3.7 the function is Y = 2X as each item sells for \$2, so fifty units equates to \$100 in sales.

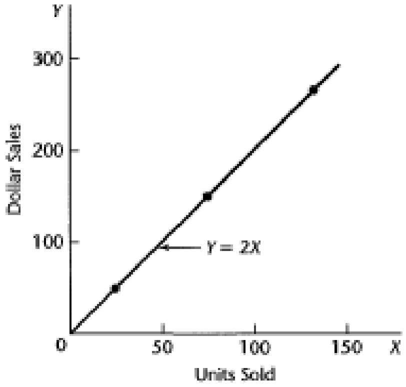


Figure 3.8 Simple Linear Regression example
Source: (Neter et al., 1996)

However, most if not all Linear Regression models will have multiple variables and will not exhibit a linear relationship such as the one in Figure 3.8, where all data points lie exactly on the function line.

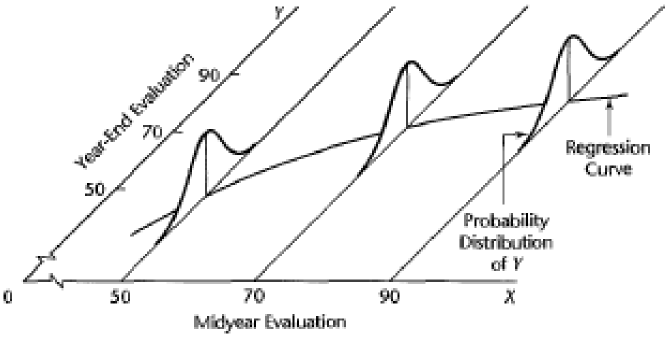


Figure 3.9 Linear Regression Curve
Source: (Neter et al., 1996)

Figure 3.9 illustrates a more typical Linear Regression model, where the function line is a curve, as opposed to a straight line. The data points will not necessarily lie directly on the regression curve, but will most likely be spread around the curve, but should fall within the probability distribution of Y, and hence be close to the curve (Neter et al., 1996).

Linear Regression and other variants of Regression such as Logistic or Probit Regression models would have one of the most popular forms of modelling techniques for mortgage prediction models in the past, but in more recent years other approaches such as Neural Networks and Support Vector Machines have taken over.

(Utrilla and Constantinou, 2010) proposed a Logistic Transition Matrix Approach (LTMA) to modelling mortgage default. The LTMA model looks at each individual borrower and works out the probability of the mortgage moving from one status to another. If the mortgage is up to date with all payments, then status of the mortgage can change from (1) current, to (2) 30 days in arrears or (3) paid off. If a mortgage is already 30 days in arrears then there are a number of other statuses it can change to including 60 days in arrears, or it can go back to being current with the repayments.

In order for the LTMA model to work correctly, it requires individual loan level data for all mortgages from the origination of the mortgage. The model works by calculating the probability of each mortgage moving from its current status to any other status, as well as working out the probability of the mortgage staying in its current status. The model generated by (Utrilla and Constantinou, 2010) performed well, as can be illustrated by Figure 3.11.

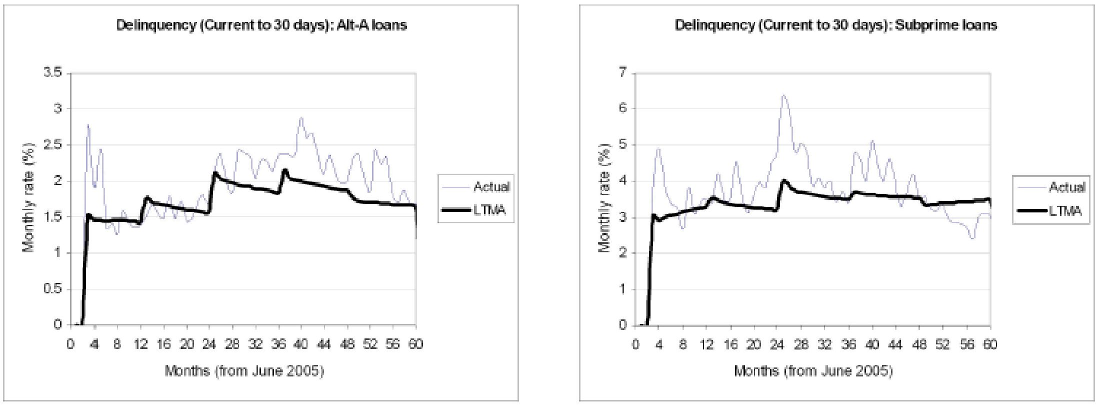


Figure 3.10 Logistic Transition Matrix Approach performance
Source: by (Utrilla and Constantinou, 2010)

(Srinivasan et al., 2011) propose a Partial Least Squares (PLS) Regression model for predicting loan defaults. The PLS model was coupled with Variable Influence on Projection (VIP) scores, to decide which data items were of most importance in the regression model, to reduce the complexity and computational ask on the model. The dataset used was imbalanced with a large number of records that did not default, so to address this they used a number of PLS models. The data was separated in to a three distinct groups or clusters, based on the predicted risk, so that the records were sorted in to Low, Medium and High risk categories. By dividing the data in to these groups the resultant models were able to achieve a better score as the loans tended to naturally group together in to those categories.

3.4.4 Support Vector Machines

Neural Networks were built in large part due to the inspiration provided by the human brain and how the design of the brain could be used in computers, but Support Vector Machines came about largely as a result of the theory behind the Statistical Learning Theory (Vapnik, 1999). Support Vector Machines (SVMs) compare favourably to many other algorithms and can perform very well in varied fields such as pattern recognition, medical diagnosis and text categorisation (Shin et al., 2005).

SVMs seek to be able to separate the data in the model in to separate classes using a linear Hyperplane (Vapnik, 1999). In Figure 3.10 the two distinct classes are linearly separable by the Hyperplane H, with a number of data points from each class sitting on the planes H_1 and H_2 respectively. The Hyperplane is chosen as the line that gives the maximum distance between the two classes and linearly separates them. The data points for each of the classes in Figure 3.12 that sit on the respective planes H_1 and H_2 respectively are known as the support vectors, as these are the data points that will decided where the optimal Hyperplane will exist (Shin et al., 2005).

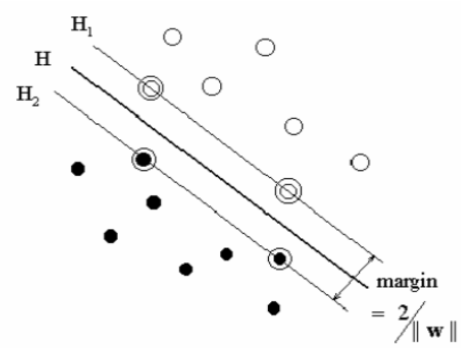


Figure 3.11 Support Vector Machine separating classes
Source: (Wang et al., 2007)

For many datasets the classes cannot be easily linearly separated, so in this case the SVM will seek to map the inputs to a high dimensional feature space using non-linear functions. Within this high dimensional feature space, the optimal Hyperplane should be able to linearly separate the data (Vapnik, 1999). Even in the high dimensional feature space it is not always possible to correctly classify each data item, but the distance from the item to the Hyperplane will be taken as a measure of the error.

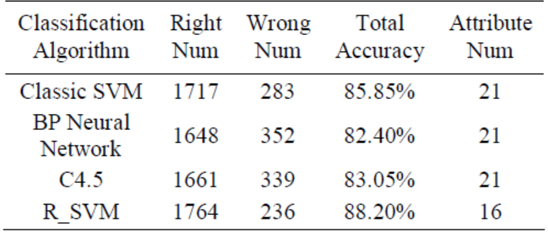
SVM models are broadly used in bankruptcy predictive models with both (Chaudhuri and De, 2011) and (Shin et al., 2005) finding them to perform better than Neural Networks at predicting bankruptcy. (Sun and Li, 2012) investigated the application of an ensemble of SVMs in bankruptcy prediction and found that an ensemble with a number of base SVM classifiers performed considerably better than individual SVMs were capable of doing.

(Bellotti and Crook, 2012) tested the performance of SVMs against more traditional default models including Logistic Regression (LR), Linear Discriminant Analysis (LDA) and K Nearest Neighbours (KNN). They found that the SVMs in general performed better than the other methods, and that the SVMs were also useful for extracting the most important attributes associated with default risk.

(Wang et al., 2007) proposed a methodology of using Rough Set (RS) as a feature selection algorithm to reduce the number of dimensions in a Chinese mortgage dataset. The reduced dataset was then fed in to an SVM and the performance of this SVM was compared to that of an SVM, a Back Propagation Neural Network and a Decision Tree that did not use RS for feature selection. The performance of the three models that did not use RS for feature selection were all quite favourable, though the SVM model combined with the feature selection of RS performed considerably better than the other three as can be seen from Table 3.2. The overall accuracy rate of the SVM model with RS was 88.20%, though the research does not mention how well the minority class is classified.

Table 3.2 Rough Set SVM comparison

Source: (Wang et al., 2007)



Classification Algorithm	Right Num	Wrong Num	Total Accuracy	Attribute Num
Classic SVM	1717	283	85.85%	21
BP Neural Network	1648	352	82.40%	21
C4.5	1661	339	83.05%	21
R_SVM	1764	236	88.20%	16

3.4.5 Boost Models

A Boost predictive modelling algorithm seeks to harness the power of a number of weak classifiers, and turn the output from these classifiers in to a classifier with a much better classification score. The reasoning behind boosting is that it is much easier to find a number of rules that perform moderately well, as opposed to finding the one rule that performs exceptionally well (Schapire, 2003).

The weak classifiers are repeatedly fed different subsets of the training data, and this in gives a classification score for each of the iterations. From the classifications output of the weak classifiers, the boosting algorithm assigns a weight to the scores of each of the classifiers. The algorithm focuses on the examples that have been misclassified the most, and disregards the examples that have been classified correctly so that the algorithm is geared towards correcting these errors (Schapire, 2003).

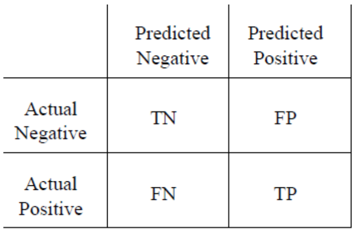
The iterative approach achieves a better classification outcome for each of the weak predictors, and then the overall classification model is built by taking a weighted majority vote from each of the classifiers (Schapire, 2003). This weighted majority vote gives a model with a much stronger predictive capability when compared to the individual weak predictive models.

One of the most commonly used boosting methods is the Adaptive Boosting algorithm (AdaBoost). The AdaBoost algorithm works by adjusting the boosting during each iteration depending on the errors of the classifiers (Freund and Schapire, 1995). In this way the algorithm adapts to the outputs of the model, and seeks to address the incorrectly predicted outcomes. A more recent boosting algorithm was proposed by (Finlay, 2008) called Error Trimmed Boosting or ET Boosting. The ET Boosting algorithm is different from AdaBoost as it does not apply weights to the observations, and each observation is equally likely to be included in the construction of a classifier (Finlay, 2008)

Both AdaBoost and ET Boost were applied to credit and loan datasets by Finlay in order to test the performance of the newly proposed ET Boost methodology. ET Boost performed more favourably than AdaBoost in all cases, though further investigation is required to ensure that the algorithm can perform well in any situation.

3.5 Model Evaluation Methods

In the same way that there are many ways and techniques for building predictive models, there are many different ways to assess these models. One of the main methods used to be able to quantify the accuracy of a model is the overall accuracy rate. The accuracy rate is defined as the number of outcomes that were correctly predicted, over the total number of outcomes in the dataset.



	Predicted Negative	Predicted Positive
Actual Negative	TN	FP
Actual Positive	FN	TP

Figure 3.12 Confusion Matrix
Source: (Chawla et al., 2011)

Figure 3.13 shows a confusion matrix that is generated for any classification or prediction models, and it shows the number of records that were classified correctly as either true positives (TP) or true negatives (TN), as well as the records that were misclassified as either false negatives (FN) or false positives (FP). Using the Confusion Matrix as a guide, the accuracy rate can be defined as follows:

$$\text{Accuracy} = \frac{TP+TN}{TN+FP+FN+TP}$$

The inverse of the accuracy rate is the misclassification rate, which is defined as the number of false negatives and the number of false positives divided by the total number of records in the dataset. If there are one hundred items in a dataset and ninety of these are classified correctly as either TP or TN and the other ten are misclassified, then the accuracy rate of the model will be 90%, while the misclassification rate will be 10%.

Both the accuracy rate and misclassification rate can be misleading, as they tend to overlook deficiencies in datasets that are imbalanced. If in the example of the dataset given previously, there were ninety items correctly classified, but all of these ninety belonged to one class, and the ten that were classified incorrectly all belonged to the opposite class, then the model would not actually have performed to 90% accuracy as the accuracy rate would suggest. For imbalanced datasets very often the focus is on the prediction rate for the minority class, as the minority class is usually the class of interest. The overall accuracy rate is not a true indicator of how a model has performed if there is a minority class that is under-represented in the model, so other methods are required to measure how good a model actually is.

Recall is the rate of TP divided by the total number of TP + FN, and this measure is also known as Hit Rate or Sensitivity. This measure is a more accurate reflection of the true accuracy of the model, where the TP class is the class which is the most important in the prediction. The recall rate gives a better reflection of how the positive target class is being classified, which is a good indicator of how the predictor is performing, especially in the case of imbalanced datasets. The Precision value of the model is the number of TP divided by the total number of TP + FP. Recall can be viewed as the accuracy rate for predicting all of the positive class records, whereas precision can be viewed as the accuracy of all the records that have been classified as positive by the model. Specificity is another method of evaluating the performance of a model, but working out how well the model has performed at classifying the negative class. Specificity is defined as a measure of how well the model has predicted all negative class examples.

$$\text{Recall} = \frac{TP}{TP+FN} \qquad \text{Precision} = \frac{TP}{TP+FP} \qquad \text{Specificity} = \frac{TN}{TN+FP}$$

There are many different approaches to modelling data classification problems, but it can be hard to compare and contrast multiple different types of classifiers against each other. Different models produce different outputs and metrics, and hence it can prove difficult to be able to decide which model is the best one to choose.

Testing against unseen data

Predictive models are normally tested against data that the model has not previously seen, to work out the true predictive capability of the model. The misclassification rate, recall rate, precision and ROC curves are normally generated from the training and validation datasets, but the model has been trained and validated using this data. By testing the model against previously unseen data, a true reflection of how the model performs and generalises on data it has not previously seen can be achieved. Very often a model may be very accurate when it is being tested against both the training and validation datasets, however when tested against previously unseen data it may not perform quite as well. The model may become over familiar with the data in the training and validation datasets thus it is not able to generalise well when faced with a different dataset. This is also called overfitting of the model and is to be avoided when building a model, as it will bring the accuracy of the model down considerably.

Normally when building a new model the overall data available for the model will be split in to two datasets, with a small dataset being held out from the model while the larger dataset is split in to training and validation datasets to build and test the model. The hold-out dataset may simply be data for the same time period, or it may be data for a different calendar month.

(Scheurmann and Matthews, 2005) tested their ensemble model against previously unseen data and achieved comparable results to the training and validation datasets used to be train and test the model. This would indicate that their model while being relatively accurate, is able to generalise well for data that it has not previously seen.

3.6 Conclusions

This chapter has presented a discussion on the available literature for predictive modelling, and has focused on predictive models for the classification of mortgage arrears and default. Different modelling techniques were introduced including Neural Networks, Support Vector Machines and Regression Models. Each of the different types of modelling techniques have their strengths as well as their weaknesses, and there is not a simple “one model fits all” scenario where the same algorithm can be used for all datasets. As a result it is generally necessary to try a number of different approaches for each classification or prediction problem. Much of the current literature on mortgage arrears would suggest that Neural Networks are the most popular of all the algorithms used to model arrears and default. This does not suggest though that Neural Networks will likely produce a good result for all datasets.

The chapter also presented some of the literature on working with imbalanced datasets, and how best to deal with them. There are numerous different methods utilised for dealing with imbalanced datasets including over and under sampling as well as the creation of synthetic data using a number of different algorithms to decide what data is to be sampled and created. Some of these methods will likely be implemented in this research project to address the imbalance that currently exists in the dataset.

4. EXPERIMENT DESIGN

4.1 Introduction

This chapter will present the design of the experiment to be carried out as part of the research project. It will cover in some detail the data available for the model, as well as any transformations or pre-processing that has been applied to the data in order to have it in the correct format for the modelling process.

Details of what algorithms and processes are to be implemented as part of the modelling phase, as well as how the experiment is to be carried out will be discussed. As the dataset available for the model has a relatively low instance of mortgage arrears, a number of methods of trying to address the imbalance in the dataset will be considered, and these approaches will be covered in respect to this experiment.

The evaluation methods to be used to assess the individual algorithms and models to be implemented will be discussed in detail, as these will be used to decide which model is to be chosen as the best model for the prediction engine.

4.2 Focus of experiment

The focus of the experiment is to design, build and test a predictive model that will be able to predict what mortgages are likely to go in to arrears with their payments, and potentially default on their credit obligations.

The main aim of the experiment is to test the predictive capabilities of customer spend analysis and negative equity estimates in a mortgage arrears prediction model. This will be achieved by deriving customer spending habits and trend analysis from all of the customer's transactions over the previous eighteen months, where available.

Once the predictive model has been built and tested on current data, a further aim of the experiment is to work out what is the optimal time for predicting arrears. It is suggested that in order to be able to help out any customers in distress on their mortgage, that it would be useful to know in advance what customers are likely to be in difficulty, so that they could be contacted by the lender in order to be able to work out a solution that would be beneficial to both parties.

The model to be created as part of this experiment should be able to accurately predict what customers are likely to go in to arrears on their mortgage payments. However, if this could be accurately predicted a number of months in advance of the arrears event, then this would be more valuable than a model that would predict what customers were going to go in to arrears in the following month. It is likely that there will be a trade-off in the accuracy of the model if it is being run a number of months in advance of the event, but this trade-off can be offset with the knowledge that most of the customers in difficulty have been predicted.

It would be suggested that if a lender knows in advance that a borrower is going to be in difficulty, then they can address the situation in a timely manner and ultimately potentially avoid the borrower getting in to difficulty. Therefore, the potential drop in accuracy could be offset by the reduction in the number of customers who actually experience arrears.

4.3 Data

The vast majority of the data required for the experiment already exists on a Teradata warehouse within Lender A. Therefore, most of the data required for the modelling should be easily accessible. In some cases the data will need to be derived for the model, and this is certainly the case with the derived spending habits from the transactional data for each mortgage.

The different data items required for the predictive model will be sourced from a number of different entities from within the data warehouse including the Customer, Application, Mortgage, Transaction and Account entities. Each of these entities are generally fed from separate source systems, that exist as separate entities outside of the warehouse with their own rules and integrity constraints. As a result there may be some referential integrity issues with the data, but for the purposes of this experiment only data that meets all existing referential criteria within the warehouse will be included in the data utilised. Any data for whatever reason is believed to be incorrect in terms of referential integrity will not be included in any model generation.

For the purposes of this experiment, each mortgage will be taken in its entirety. In some cases a mortgage may have multiple loan accounts associated with it, as a mortgage may be split across a number of different products or a borrower may have received a top-up loan or released equity from the property. A classic example of this is where a customer has a mortgage for €350,000 with the mortgage split across two loan accounts, one on a fixed rate, with the other portion on a tracker or variable rate. For the purposes of this experiment, this will be treated as one mortgage, though in some cases this may be reported as two mortgages, especially when reporting on mortgage types. By rolling up the mortgage accounts associated with each property it will be ensured that each property will only be included once in the analysis.

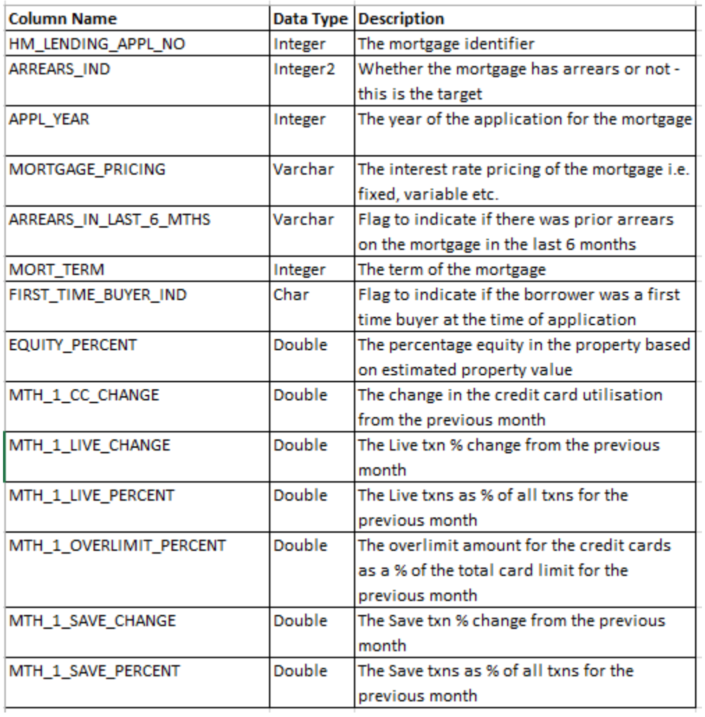
A mortgage must have at least one customer associated with it, though there can be multiple customers associated with one mortgage. The most common number of customers associated with a mortgage is two, though mortgages can exist with three or four customers associated with them. For the purposes of the data items that are at a customer level any of the numeric variables such as salary will be aggregated, so that the mortgage total will be a sum of all the constituent customers associated with the mortgage. An example of this would be the total salary for each customer. If there are multiple customers associated with a mortgage, then the mortgage level salary will be an aggregation of the individual customer salaries. For other values such as the age of the customers associated with the mortgage the minimum, maximum as well as the mean values for all associated customers will be used.

Similarly for data items that are currently held at account level these will be rolled up to mortgage level, so that the mortgage representation will be a summation of all the values associated with the constituent accounts. Some examples of this include the balance on the mortgage as well as the outstanding arrears amount where applicable.

4.3.1 Data items to be included

Much of the data to be included in the dataset for the model will be coming directly from the mortgage entity on the data warehouse. Some of the data items taken from the mortgage entity will include the mortgage term, the outstanding balance of the loan associated with the mortgage, the property type and location of the property. Some of the customer level information (which will be aggregated up to each mortgage) will be whether or not the borrowers have any dependent children, the derived salary and age of the borrowers. The account level information will include the interest rate type of each of the loans associated with the mortgage, the dates that the accounts opened as well as the current balance on the loans. Table 4.1 details examples of some of the data items that will be included in the model. The full listing of all data items included can be found in Appendix A.

Table 4.1 Example data items to be included in the model



Column Name	Data Type	Description
HM_LENDING_APPL_NO	Integer	The mortgage identifier
ARREARS_IND	Integer2	Whether the mortgage has arrears or not - this is the target
APPL_YEAR	Integer	The year of the application for the mortgage
MORTGAGE_PRICING	Varchar	The interest rate pricing of the mortgage i.e. fixed, variable etc.
ARREARS_IN_LAST_6_MTHS	Varchar	Flag to indicate if there was prior arrears on the mortgage in the last 6 months
MORT_TERM	Integer	The term of the mortgage
FIRST_TIME_BUYER_IND	Char	Flag to indicate if the borrower was a first time buyer at the time of application
EQUITY_PERCENT	Double	The percentage equity in the property based on estimated property value
MTH_1_CC_CHANGE	Double	The change in the credit card utilisation from the previous month
MTH_1_LIVE_CHANGE	Double	The Live txn % change from the previous month
MTH_1_LIVE_PERCENT	Double	The Live txns as % of all txns for the previous month
MTH_1_OVERLIMIT_PERCENT	Double	The overlimit amount for the credit cards as a % of the total card limit for the previous month
MTH_1_SAVE_CHANGE	Double	The Save txn % change from the previous month
MTH_1_SAVE_PERCENT	Double	The Save txns as % of all txns for the previous month

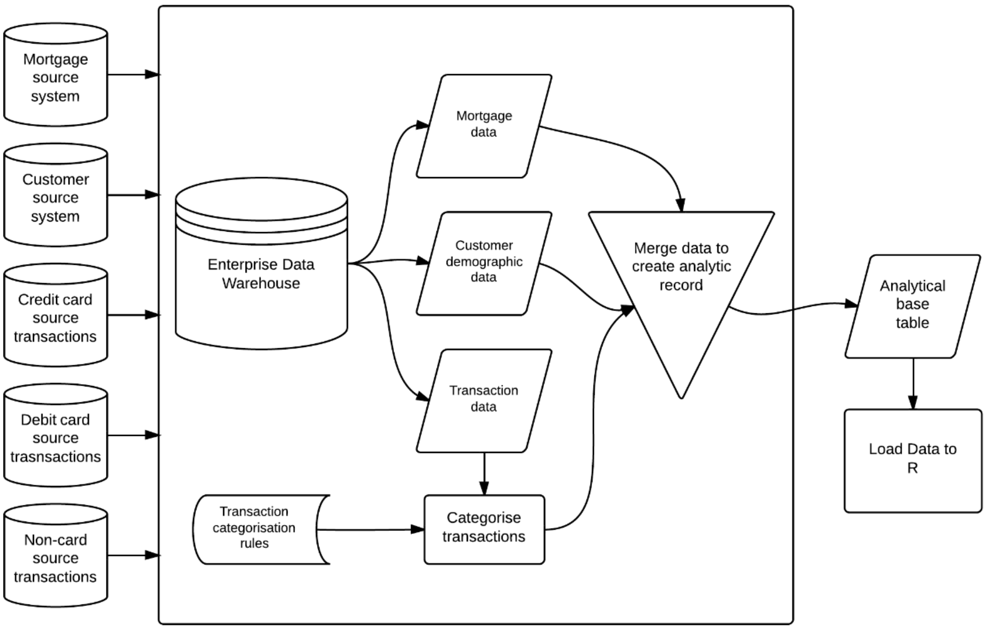
For each mortgage included in the experiment all available data items will be selected and included in the model. The dataset will be created by first identifying what mortgages are in scope for both the training/validation dataset and the holdout dataset. The training/validation dataset will be the taken from a periodic table with a snapshot of all the mortgages at the end of September 2013, while the holdout dataset will be generated from a snapshot of the mortgages at the end of February 2014. The majority of mortgages currently on the books would have been present at the end of September 2013 as well as February 2014, so care must be taken they do not end up in more than one dataset. This will be achieved by ensuring that when the holdout dataset is created, none of the mortgages in the dataset will be included in the training/validation dataset.

4.3.2 Data Collection and Preparation

The data will be collected from the data warehouse for all mortgages in scope for the experiment, and will be split up in to the training/validation data and the holdout testing data. The data to be used for building the model and for training and validation will be a sample of mortgages taken from a periodic snapshot of mortgages at the end of September 2013. The holdout dataset will contain the same fields but will be taken from a periodic snapshot of the mortgage data at the end of February 2014.

The code used to extract the two separate datasets from September 2013 and February 2014 will be the exact same to ensure there are no quality issues with the data. Each subset of mortgages will then be used to build the training and validation datasets, and the February dataset will be used to build the holdout dataset. The total number of mortgages in arrears is 7,525 mortgages, and this represents roughly 10.65% of all the mortgages available for the experiment. Both the holdout testing data set and the training and validation dataset will have the same proportion of arrears to non-arrears, so that the model will be classifying records in a realistic way. As discussed later in this chapter, a number of measures will be implemented to try to address the class imbalance in the dataset, but this will only be for the training/validation dataset, as the holdout testing dataset and any real datasets the model will be run against will have this class imbalance.

The two datasets (for Sep 2013 and Feb 2014) will be run in parallel to return all mortgages at each point in time, and the associated information. As the vast majority of mortgages in the September dataset will also be in the February dataset the training/validation and holdout datasets will have to be built in such a way to ensure that the same mortgage(s) are not in both datasets. This will be achieved by firstly building the holdout dataset by randomly selecting a number of mortgages that are in arrears, so that the ratio of mortgages in arrears/non-arrears in both the holdout dataset and training/validation datasets will be the same. Subsequently a random sample of mortgages not in arrears will be added to the holdout dataset, so that the ratio of arrears to non-arrears is the same as in the overall dataset. Then the training/validation dataset will be built by selecting the remaining mortgages in the September dataset in arrears, that are not already include in the holdout dataset. Consequently all other mortgages in the September dataset not in arrears that have not been included in the holdout dataset are added to the training/validation dataset.



Mortgage source system
Customer source system
Credit card source transactions
Debit card source trasnsactions
Non-card source transactions
Enterprise Data Warehouse
Mortgage data
Customer demographic data
Transaction data
Transaction categorisation rules
Categorise transactions
Merge data to create analytic record
Analytical base table
Load Data to R

Figure 4.1 Flowchart of data for modelling

At this stage all available mortgages will be used in either the training and validation dataset or the holdout dataset. A table will be created that holds each mortgage identifier, and which dataset it is included in so that there will be no confusion at a later stage. The final datasets for each the training/validation and holdout datasets will be held separately, but for the purposes of generating all of the transactions for the accounts associated with the customers, both sets of data will be combined in to one table, along with the field that indicates which dataset the mortgage will belong to.

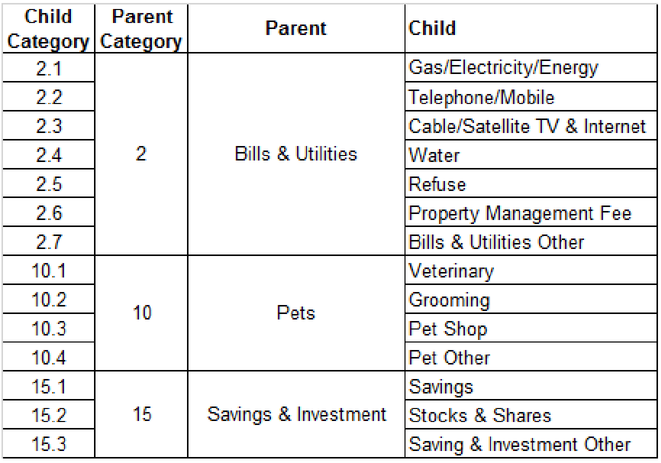
Once it has been decided what data items are to be included in each of the two datasets, the next step will be to gather all of the transactions for the accounts associated with each of the mortgages. Transactions for all accounts associated with each of the mortgages will be collected from the four main transaction tables (Credit Card, Laser Transaction, Visa Debit and Transaction Detail for all accounts for eighteen months. If a mortgage is in arrears at the time the data is extracted for the experiment, then the eighteen months of transactions will be taken for the months immediately prior to the arrears. The month where the mortgage started to go in to arrears will not be included in the transactions as the borrower would not have fulfilled their mortgage obligations for this month as the previous month would have been the last full payment the borrower made on the mortgage.

For all other mortgages that are not in arrears the transactions will be taken for the eighteen months immediately prior to the date from which the mortgage was selected – September 2013 or February 2014. This means that for the September 2013 mortgages the transactions will be up to and including August 2013, and for the February 2014 mortgages the transactions will be up to and including January 2014. These transactions will be employed in the model, but they will not be fed directly in to the model in their raw format. Therefore pre-processing will be required to get the transactions in to a usable format.

4.3.3 Processing the Transactional Data

The transactional data that will be collected as part of the experiment will be all transactions for customers associated with the mortgages in the sample. Each of the transactions will need to be identified and captured, as well as categorised based on the rules for the Money Manager Application categorisation.

Table 4.2 Transaction Parent/Child categories



Child Category	Parent Category	Parent	Child
2.1	2	Bills & Utilities	Gas/Electricity/Energy
2.2			Telephone/Mobile
2.3			Cable/Satellite TV & Internet
2.4			Water
2.5			Refuse
2.6			Property Management Fee
2.7			Bills & Utilities Other
10.1	10	Pets	Veterinary
10.2			Grooming
10.3			Pet Shop
10.4			Pet Other
15.1	15	Savings & Investment	Savings
15.2			Stocks & Shares
15.3			Saving & Investment Other

Lender A offers the Money Manager Application to its personal customers to allow them to keep track of their spending. It automatically categorises all of the transactions for a customer’s accounts, and allows the customer to view the aggregated category totals through an application dashboard, where they can keep track of their spending. There are a total of sixteen parent transaction categories that are each further split down in to a number of child categories. Examples of some of these parent/child categories can be seen in Table 4.2.

For all transactions that have taken place on a Credit or Debit card the transactions are categorised using the Merchant Category Code (MCC), which defines what type of purchase or service the customer availed of. For other transactions that did not take place on a card, there are a set of rules defined that will decide what category the transactions should belong to. These rules are a combination of the source of the transaction as well as the narrative and amount. At present the categorisation and storage of these categories takes place in the database of a third party supplier, and the information is not fed back in to the data warehouse. As a result the rules have been replicated on the data warehouse, and all transactions will be categorised here.

4.3.4 Spend/Save/Live/Mortgage Categorisation

The idea of grouping transactions in to Spend/Save/Live high level groups was adopted from the American bank Moven[1], who are an internet only bank in the US. They offer a simple banking service that supplies a checking account (current account) with a debit card, which has no fees for the customer. The also provide an online banking and mobile app service that allows the customer to easily keep track of their spending throughout the month. The app categorises all of the customer's transactions in to Spend/Save/Live and keeps the customer updated throughout the month on how much they have spent in each of the categories.

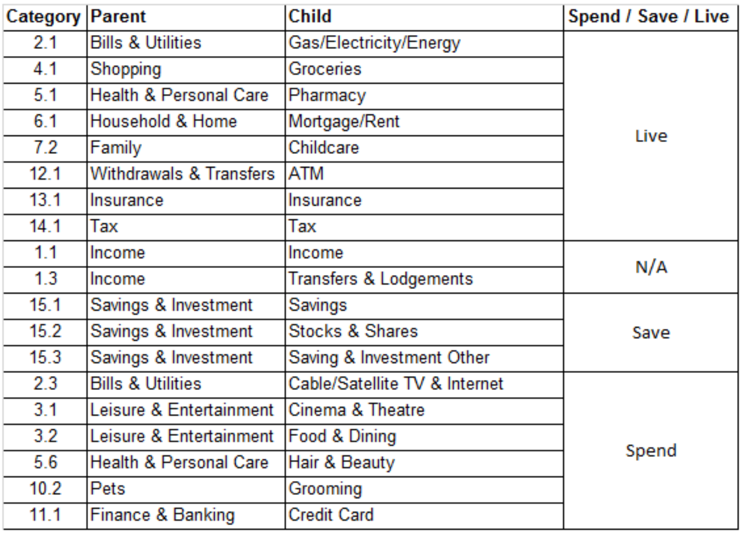
The Spend category is seen as the category for discretionary spending which will include leisure activities such as eating out and purchasing personal electronics etc. The Save category details all the saving the customer has done throughout the month and the Live category tracks spending on the essentials required for a customer to live such as bills, mortgage/rent and food shopping. For this particular analysis it is thought that the mortgage payments might be better suited to be in their own category, so they are put in to a category called Mortgage. It is believed that by taking the existing Money Manager

[1] https://www.moven.com/

categories Lender A has, it should be able to use the Spend/Save/Live/Mortgage high level categories to gain insight in the area of mortgage arrears. At an anecdotal level it would be supposed that for any borrower who is getting in to difficulty the level of discretionary spend would drop at a noticeable level, while the live category would stay somewhat static, and with savings likely to fall when borrowers are in further difficulty, when money earmarked for savings may be used to help with mortgage payments.

Once all of the transactions have been categorised in to the individual Money Manager categories the transactions need to be aggregated up in to the Spend/Save/Live/Mortgage categories. This is achieved by assigning either Spend, Save or Live to all of the Money Manager categories in a reference table where applicable, so that the transactions can be aggregated up to each of the high level groupings. For the purposes of this experiment only transactions where the customer has debited their account are going to be used in the derivation of fields for the model. There are some transactions that cannot be placed in to any of the three groups, but most of the Money Manager categories can be. Credit transactions do not naturally fit in to any of these categories, but they are not going to be used for the modelling anyway. Examples of the Money Manager categories and their corresponding Spend/Save/Live categories can be found in Table 4.3.

Table 4.3 Spend/Save/Live Categorisation



Category	Parent	Child	Spend / Save / Live
2.1	Bills & Utilities	Gas/Electricity/Energy	Live
4.1	Shopping	Groceries	
5.1	Health & Personal Care	Pharmacy	
6.1	Household & Home	Mortgage/Rent	
7.2	Family	Childcare	
12.1	Withdrawals & Transfers	ATM	
13.1	Insurance	Insurance	
14.1	Tax	Tax	
1.1	Income	Income	N/A
1.3	Income	Transfers & Lodgements	
15.1	Savings & Investment	Savings	Save
15.2	Savings & Investment	Stocks & Shares	
15.3	Savings & Investment	Saving & Investment Other	
2.3	Bills & Utilities	Cable/Satellite TV & Internet	Spend
3.1	Leisure & Entertainment	Cinema & Theatre	
3.2	Leisure & Entertainment	Food & Dining	
5.6	Health & Personal Care	Hair & Beauty	
10.2	Pets	Grooming	
11.1	Finance & Banking	Credit Card	

Much of the transactional data to be used for the experiment will seek to see trends and patterns in the spending habits of the borrowers associated with the mortgages. It is for this reason the transactional data will be taken for eighteen months, so that the activity of the borrowers associated with the mortgage can be viewed over this time period. It is not expected that the changes in spending will be as evident a number of months in advance of the mortgage going in to arrears, but this should become clearer during the execution of the model. When the model is being built, six months of transactions will be used for any iteration of the model, so the eighteen months of transactions will allow the modelling to be completed for mortgages up to twelve months in advance of arrears. The Spend, Save and Live percentage split will be calculated for each mortgage for each month as a total of all of the debit transactions, and the difference between a month and the preceding month will also be calculated as a percentage change in terms of the value of the transactions.

In addition to deriving the transactional data for all mortgages for eighteen months there will be a number of other numeric values that will be captured for each mortgage over the 18 month period. Some of the values include the derived salary, total savings balance, credit card balances and mortgage repayments. These values will be calculated for each of the 18 months in a similar way to the transaction level data, and the percentage change month on month will also be calculated for each of the fields.

4.4 Software used

The experiment will be designed and implemented in R and Rattle. R is a statistical language and Rattle provides a data mining Graphical User Interface (GUI) that sits on top of R. Rattle allows the user to model a dataset without having to enter the specific R code required. The GUI does provide all R code in the log file, so that the code can be re-used outside of the GUI.

Rattle offers rich functionality for importing, exploring and modelling datasets. The capability exists to create Decision Trees, Random Forests, Boost Models, Support Vector Machines (SVM), Regression models, Neural Networks and Survival models. For most of these modelling techniques the input parameters can be tweaked depending on the situation and the modelling requirements.

For the most part the parameters used for the models will tend to be the default settings, though for the SVM technique, there are multiple algorithms that can be used to build the model. In some cases tweaking the settings of a model can tend to lead to overfitting of the training and validation data in order to improve the accuracy of the model. The settings will only be changed for the modelling techniques when absolutely necessary.

4.5 Building the Models

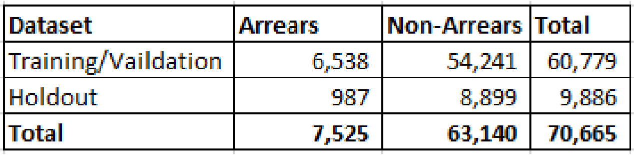
Each of the models produced will be trained using the training data available. The models will individually decide which data items are most important as part of the training phase, and then the validation dataset will be used to validate the model output.

In order to have a baseline prediction model to test against, a simple model will be built, without any of the transactional spend or the derived negative equity and LTV data. The scores achieved by this model will allow a comparison to be generated against the models that are created using the newly derived negative equity and transactional spend data. It is expected that creating a model with the additional data available from the transactional spend and the derived equity and LTV to value rates it should be possible to increase the accuracy of the predictive models. By comparing the results of this basic model with the newly created models it should be apparent if the new data has had an impact increasing the predictive capability and accuracy of the models.

For the purpose of building, training and testing the model the data exported from the Data Warehouse will be split in to a number of different partitions. The bulk of the data will be used for building and training the model, with a further set of the data used for validation and the holdout dataset that will be used to test the predictive capability of the model against previously unseen data. The final sample of mortgages available for the design, build and testing phase of the model is 70,665 mortgages. The sample size available for the training and validation of the model is 60,779 with the holdout sample containing the other 9,886 records. In the training/validation dataset there are 6,538 mortgages in arrears, and in the holdout testing dataset the number of arrears is 987.

The 60,779 records that are to be used for the training and validation dataset will be split using a ratio of 70/30, so that 70% of the data will be used for training with 30% being used for the validation dataset. The data will be split so that the training and validation data has the same ratio of arrears to non-arrears which will allow the model to be consistent across both datasets.

Table 4.4 Breakdown of arrears/non-arrears in all datasets



Dataset	Arrears	Non-Arrears	Total
Training/Vaildation	6,538	54,241	60,779
Holdout	987	8,899	9,886
Total	**7,525**	**63,140**	**70,665**

The ratio of arrears to non-arrears in both the training/validation and holdout datasets are very similar, with the ratio being 10.76% in the training/validation set and 9.98% in the holdout dataset. As a result the model will have comparative levels of arrears and non-arrears in each of the datasets which will allow for consistency.

The four different types of models to be created as part of this experiment are Decision Trees, AdaBoost, Support Vector Machines and Neural Networks. These are some of the most commonly used algorithms, and they should all give relatively good scores for this dataset. There are many other types of models that could be used for this process, but these have been chosen as they are readily available in both R and Rattle, and they are well known and universally recognised.

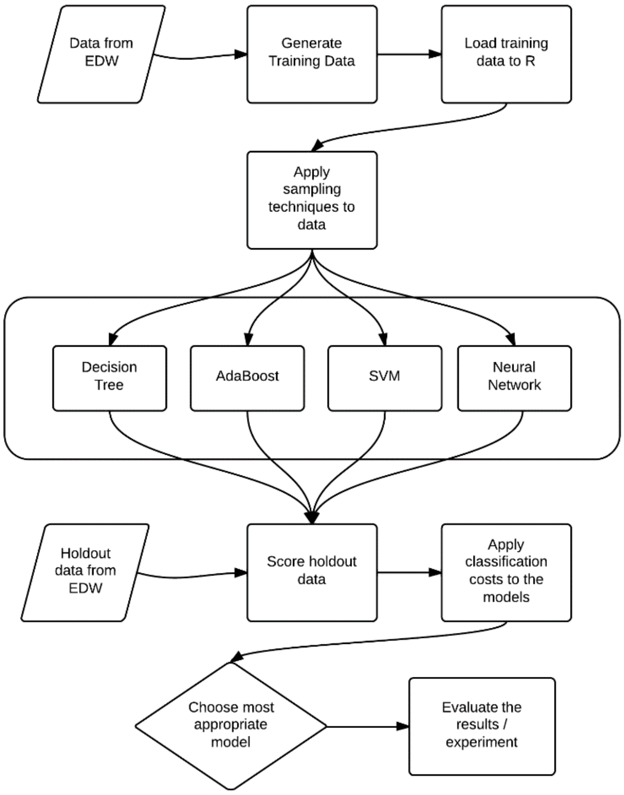

Data from EDW
Generate Training Data
Load training data to R
Apply sampling techniques to data
Decision Tree
AdaBoost
SVM
Neural Network
Holdout data from EDW
Score holdout data
Apply classification costs to the models
Choose most appropriate model
Evaluate the results / experiment

Figure 4.2 Overview of model experiment process

4.5.1 Addressing the class imbalance in the dataset

The dataset of mortgages available for the design and implementation of this experiment is heavily weighted towards mortgages that are not in arrears, as would be expected in any normal mortgage book. However this poses a problem for building a predictive model, as most models require a dataset to be balanced. If the dataset is heavily weighted towards one class, then the algorithm will likely predict this class very well, to the detriment of predicting the minority class. A number of methods will be implemented to see if this imbalance can be addressed, so that the model will accurately predict the minority class including the following:

- Random undersampling
- Random oversampling
- Synthetic oversampling

Random Undersampling

The majority class data in the training and validation data will be randomly undersampled, so that the number of mortgages in the dataset in arrears will match the number of mortgages in the dataset not in arrears. This will be achieved by taking a random sample of the majority class equal in number to the minority class, which in the training and validation dataset is 6,538. Therefore there will be an equal split in the arrears and non-arrears in the dataset.

It would be expected that the model with the random undersampling should perform better on the predictions for the minority class than the model with the full dataset. In order to test the validity of the random undersampling method a number of samples will be tested, and the average score taken for each of the models. For each undersampling iteration the sample of majority classes will be created from within the data warehouse using the standard sampling method in Teradata SQL, where all records will be sampled without replacement, so no majority class records will exist in more than one sample.

Random Oversampling

The minority class data in the dataset will be oversampled randomly, to bring the number of records in the dataset in the minority class up to that which is equal to the majority data class. As there are so few minority class records compared to the majority class records, it would be necessary to oversample the data roughly nine times in order to balance up the two classes. This oversampling of the data while it may increase the accuracy for the training and validation datasets minority class prediction would likely see a large decrease in the holdout dataset testing due to the model being overfitted to the training and validation data.

Synthetic Sampling

Synthetic sampling will be utilised to increase the number of minority class examples in the dataset by creating sufficient new synthetic data items from the existing minority class. The new examples will be created using the SMOTE technique, whereby new items that are similar, but not the same as existing records in the dataset will be created. The SMOTE technique will be used to balance out the number of minority and majority class examples, so that there is an even number of both, and the original full number of majority samples are included in the model. The number of synthetic samples created will be the number required to bring the number of minority class examples up to the same level as the majority class.

SMOTE will be used to create the additional minority case samples in R, and then the dataset will be imported in to Rattle to be processed. It is expected that the model created using the SMOTE technique to create extra minority class examples, should perform on a similar level to the model created using majority undersampling. It’s possible that the SMOTE technique would cause the model to outperform the majority undersampling, but this may not be the case. Overfitting should be less of an issue with the SMOTE technique as the synthetic data will not be the exact same as the original data items.

4.5.2 Model Training

The known outcome for each mortgage within the training dataset will be used to train the model and to improve the accuracy of the predictions. For each record in the training/validation dataset the actual outcome (arrears/non-arrears) is known. This known outcome will be used to improve the predictions, as the models will seek to work on the records that it has misclassified. This is important for all of the models, as they use the training and validation data to both build and test the predictions. It is especially important for the AdaBoost algorithm, which iterates through the building process and weights the misclassified data items to try to correctly classify them.

During the building of the specific models the training data will be used to create each of the models and define what rules will be used to split the data and predict which class a mortgage belongs to. All of the training data will be used in this phase, and the validation dataset will then be used to validate the rules. Each of the different types of algorithms will be used to create a separate model, which can then be compared to each of the other models. By comparing how the models perform against the training, validation and holdout datasets: the best model will be chosen on overall performance.

4.5.3 Model Validation

After each of the algorithms has created a model using the training data the validation dataset will be used to evaluate the performance of the model. The training dataset will have formulated the rules for the algorithm and the validation dataset will help to appraise the performance of these rules when they are applied to a different dataset. There may not be vast differences in the training and validation datasets, but if some of the rules were built on patterns in the data of the training dataset that do not exist in the validation dataset, then this will be obvious from the performance of the model.

Ideally the model should perform to a similar level on both the training and validation datasets. If the performance of the model on the validation dataset declines drastically it is a sign that the model has been overfitted to the training data. However, just because the model performs comparatively well on both the training and validation datasets this does not mean that the model will perform well previously unseen data.

4.6 Evaluation Methods

The models produced as part of this experiment will be evaluated in a number of ways to ascertain how well they have performed. Each model will be evaluated on how it has performed using the training, validation and holdout testing datasets. It would be expected that each of the models should perform well with the training data and should also perform comparatively well with the validation data, provided no overfitting has occurred. The models will likely perform quite differently against the holdout testing data, depending on how different the holdout dataset is to the training/validation dataset and also how good each model is at generalising. The focus of a classification or prediction model is always to get as accurate a prediction as possible, but without overfitting. If the model is overfitted on the training data, then the performance of the model with unseen data would be expected to drop considerably.

4.6.1 Model Accuracy

As covered previously in Section 3.5 there are many different ways to evaluate the performance of a classification or prediction model. Some of the methods are to use the misclassification rate or the overall accuracy rate. These two values are the inverse of one another, as the misclassification rate is the percentage of records that have been misclassified, while the overall accuracy is the number of records that have been classified correctly.

In the case of classification problems where the data is imbalanced as it is in this experiment, using just the misclassification or overall accuracy rate would be misleading. This is because if most of the majority class items were classified correctly even if the minority class was completely misclassified, then the overall accuracy rate would still be quite high. While the overall model accuracy should be as high as possible there is more focus on how the model performs at classifying the minority class (the mortgages that are in arrears).

For the purposes of evaluating how the model has performed with the minority class the focus will be on the Sensitivity or Recall rate. The formula for calculating recall is as follows: $Recall = \frac{TP}{TP+FN}$. This rate will inform us of how well the model has been able to classify all of the mortgages that are in arrears. The True Positives (TP) are the mortgages actually in arrears that have been classified as being in arrears, while the False Negatives (FN) are the mortgages in arrears that the model has incorrectly classified.

Along with the Recall rate, the Precision rate will be used to work out how accurate the positive predictions have been. The formula for calculating the Precision rate is as follows: $Precision = \frac{TP}{TP+FP}$. The Precision rate tells us how accurate the model has been with its predicted positive class. The False Positives (FP) are mortgages that are not in arrears, but the model has classified them as being in arrears. It is possible that the model will incorrectly classify a large number of majority class mortgages as being in arrears and hence affect the Precision rate, but it is more likely that the model will struggle with the classification of the minority class items due to the imbalanced data.

Using a combination of the Recall and Precision the best model will be chosen at classifying the mortgages. The Recall rate does not take in to account how well the majority class has been classified, but for the purposes of this experiment these are not of as much interest as the minority class is the class of interest. The Precision rate will be an indicator of how badly the model has performed for the majority class, so this along with the Recall rate should be sufficient for gauging how successful the model is. The Precision and Recall rates will be calculated based on the confusion matrix, which will plot the expected outcomes for the mortgages against the actual outcomes for each example in the dataset. From the confusion matrix the overall accuracy of each model as well as the Negative Predictive value will also be calculated.

4.6.2 Cost of the model

In any prediction model a cost is associated with the misclassification of records. For most predictive models there is a much higher cost associated with false negatives than there is with false positives. The higher cost is as a result of the additional work or cost associated with getting the prediction wrong. If trying to predict mortgage arrears a prediction model classifies a good mortgage as one likely to have arrears then the bank is unlikely to lose any money as a result of the misclassification. The only associated costs may be contacting the customer, thought this is likely to be minimal. However, if the model predicts that a bad mortgage is actually not going to have arrears but does, then there may be very high costs associated with the misclassification. The costs associated may be in the form of additional capital requirements that were not foreseen or the actual costs of engaging with the customer after they have gone in to arrears. For true positives and true negatives there is generally no cost associated with the classification, as the model has correctly predicted the outcome for the data example. Therefore a cost will only need to be associated with false negatives and false positives.

A cost will be assigned to the misclassification of the mortgages in the models, with a much higher weighting being applied to the false negatives. The weighting applied to the false positives and false negatives will be agreed with the Arrears Support Unit within Lender A, who are tasked with dealing with borrowers who are in difficulty with their payments, or who are already in arrears. The misclassification costs will be based on the actual cost of contacting the borrowers, as well as the expected loss and possible write-off due to not capturing the arrears earlier, as well as the additional capital required for each mortgage in arrears that was not correctly classified.

4.7 Model testing with real data

Each of the models generated will be tested against live data to ascertain how accurate and predictive they are. The holdout dataset will be used as the live data, as this data will not have previously been seen by any of the models as part of the training and validation phase of the model building.

4.8 Alternate model building with Naïve Bayes Classifier

Lender A are currently carrying out a proof of concept with a new Big Data discovery tool from Teradata called Aster[2], which allows the user to carry out many statistical functions on data within a database, without having to utilise a second standalone tool.

As an alternative to the modelling already carried out as part of this experiment, a Naïve Bayes Classifier will be built in Aster to see how this classifier would perform compared to the models built in R and Rattle. The Naïve Bayes model will be trained with all the available training data, and then tested against the holdout dataset to assess the predictive capability of the model. The Naïve Bayes classifier works in a different way to how the other models work. For each record it calculates the probability that the record belongs to the positive class as well as the probability that it belongs to the negative class. The class with the higher probability is chosen as the predicted target class. This approach is similar to case based reasoning, where the model compares each record to all of the records in the dataset to work out which class each record should belong to.

4.9 Finding the optimum time to predict arrears

It is in the best interest of any lending institution to not only be able to predict with a good degree of accuracy what mortgages are going to go in to arrears, but also to be able to predict these difficulties before a borrower actually experiences them. The more

[2] http://www.asterdata.com/

notice a lending institution will have before a borrower gets in to difficulty the easier it will be for the bank to intervene and help the customer, and ultimately benefit both.

Part of this research project is to ascertain if it is possible to know well in advance if a borrower will be going into arrears. Ideally a bank would like to have a number of months to engage with a borrower prior to them going into arrears, but this is not always possible. Some borrowers can be wary or even scared of approaching their bank if they are in difficulty. If the bank could tell a number of months in advance that a borrower is likely to be in difficulty with their mortgage payments, it would be easier for the bank to approach the borrower and offer help to solve the crisis.

The trade-off with predicting arrears a number months in advance is that the model will likely not be as accurate at predicting which mortgages will go in to arrears. The model that will best predict arrears one month prior to the mortgages going in to arrears, will likely not perform quite as well a number of months in advance. It is possible that the activities or patterns in the data that allows the model to accurately predict arrears may be slightly different or even completely different when the arrears event is a number of months off. The ideal situation would be where the same model would be used in both scenarios, and the data itself would be the only element that would be different.

In order to test the model to see if it can predict arrears a number of months in advance, the model will be trained and validated in the same way as the model that will be built for predicting arrears one month in advance. The same data items and variables will be fed in to the model, but the time period that the data will relate to will be a number of months in advance of the arrears event. To test that the model works correctly and to find the optimum time to predict the arrears, the model will be tested against a number of different time periods. To validate the model to see how well it performs over the course of the different months, each time the model is tested the class predictions for all variables will be output, along with the probability associated with the class prediction.

By analysing the probabilities of the predicted class it should be possible to track how the model is able to predict arrears over time, with the assertion that as the arrears event gets closer the probability of the predicted class should be higher, so as to distinguish the best time to predict the arrears. The optimum time to predict the arrears will be when the largest number of arrears is predicted, while taking in to account the length of time prior to the arrears event.

The model to predict arrears in advance will be run using data for every second month, from the month prior to arrears, as far back as ten months in advance of the arrears. Therefore the models will be tested for the following months: month 0, month -2, month – 4, month -6, month -8 and month -10. The results of each of these iterations will be compared and contrasted, to find the optimum time to predict the arrears.

4.10 Conclusions

This chapter has presented the design of the experiment that will be used to predict what mortgages in the loan book of Lender A are likely to go in to arrears, and how the experiment is to be designed and carried out. The design of the data model to be used for the input into the predictive model was covered, along with an explanation of the variables to be included.

Also introduced in this chapter was the software that will be used to build the model, though there will be further details on this in Chapter five which will detail the actual implementation of the experiment.

The chapter documented how the model is to be evaluated, and how it is to be tested against a holdout dataset to give a true picture of its accuracy in predicting mortgage arrears when faced with a previously unseen dataset.

The penultimate section of the chapter covered how the model will be used to establish the optimum time to predict mortgage arrears, and how this will be achieved.

5. EXPERIMENT IMPLEMENTATION

5.1 Introduction

There has been much research about the rise in mortgage arrears since the start of the global financial crisis, with many researchers focusing on what has caused the rise in mortgage arrears and default. Much of the research has been looking at what has caused the decline in mortgage credit quality, and several researchers have tried to predict arrears based on the mortgage characteristics.

This chapter will present the implementation of the experiment being carried out as part of this research, to predict what mortgages are likely to go in to arrears. One of the theories being tested is that a borrower's spending habits should be a good predictor of how likely they are to get in to difficulty with their mortgage payments. In most cases the path in to arrears for most borrowers will be a gradual one, though some borrowers may suffer a more instantaneous decline in payment capacity.

All relevant design work and structuring of the modelling process will be discussed throughout the chapter as well as any methods applied to the dataset to seek to address the class imbalance in the dataset. The results achieved for the model will be discussed, though chapter six will go in to the results in more detail.

5.2 Overview

The experiment to be carried out as part of this research will seek to build a predictive model that should be capable of predicting which mortgages are likely to go in to arrears, with a high degree of accuracy. This will be achieved by taking a dataset of mortgages from Lender A and using this data to build the model. The model will be built with a set of mortgages where the outcome of the mortgage (arrears or non-arrears) is already known, and the model will use this data to build the build the algorithms behind the prediction mode. The ultimate predictive capability of the mortgage will be determined by scoring the model against unseen data in the holdout dataset.

The model will seek to test the predictive capability of non-traditional model inputs such as the current negative equity percentage for a mortgage and the spending habits of the customers associated with the mortgage. Within Lender A, it is thought that trends in customer spending could be an indicator of a customer getting in to difficulty, as spending habits are likely to change in the lead up to financial difficulty as the customer seeks to “tighten their belt”. The experiment will seek to test whether these non-traditional input fields are useful and significant in the building of a prediction model.

Once an accurate model has been built, the next objective will be to build a model that is capable of predicting a number of months in advance, which mortgages are likely to go in to arrears. This will allow the staff of Lender A to contact the borrowers involved to try to avoid a situation of them going in to arrears in the first place, or at the very least try and minimise the impact of the arrears as much as possible.

5.2.1 Processing the data

All of the data has been captured on the Teradata Data Warehouse of Lender A and will be exported to the statistical language R for modelling. Most data manipulations will be carried out in the SQL code on the data warehouse, though some manipulations may be carried out within R and Rattle to impute or calculate missing values.

5.2.2 Normalising the data items

Many of the data items to be fed in to the model are based on transaction categories, and whether or not there was a significant change in the amounts from one month to the next. Examples of these fields are the Spend/Save/Live/Mortgage categories, where the percentage value for each category is given for each of the six months before the arrears event, as well as the percentage change on month on month. The credit card and savings balance data to be fed in will be in the same format as the transactional data which will track changes in the totals from month to month, rather than including the actual values.

In some situations there are values that would be seen as outliers due to the nature of how the values are calculated. For the increases and decreases in the savings and credit card balances, the values are calculated as a percentage and in some cases these values are very high. An example would be when a customer has just fifty cent in their savings account one month but the following month they lodge €500 to the account. The new balance will be €500.50 but the percentage change month on month will be 100,000%, given the low starting amount, and the magnitude of the lodgement. The same will happen when a balance has decreased sharply from a high balance to a very low one.

In order to stop these values skewing the overall values for all other records in the dataset, the upper and lower range of values are capped at +/– 100%. This will remove the negative or potentially positive skew from these values. This process will be carried out in the SQL code used to extract the data, whereby any vales returned at over +/– 100% will be set to <-100 and >100 respectively. All other values between -100% and 100% will retain their actual value.

For the Spend/Save/Live/Mortgage categories and the breakdown of each for every month, the value for each category can only be in the range zero to one, as each category is taken as a percentage of the overall spend for that month. As a result the number of possible values for each will not be very high with the number of decimal places set to two, though the percentage change in month on month values could be similar to those of the account balances, where the increases or decreases could be very high. The percentage increases/decreases for these monthly changes will also be set at a maximum value of +/– 100%, in order to normalise the values and to lessen the impact of outliers.

There are a number of data items in the dataset that have numeric variables of type integer or double, but that need to be converted to categorical variables. Some of these fields need to be converted as the modelling process may treat them as continuous variables where it is not appropriate. Examples of the fields that require changing are the NO_BEDROOMS_CNT, APPL_YEAR and NUM_ACS fields. None of these values are true continuous variables but the modelling may treat them as such, so by converting them to categorical variables this will be avoided. The EQUITY_PERCENT and CURRENT_LTV values are calculated as Decimal values given the nature of the calculations, but there are a small number of mortgages where a missing value will not allow the calculation of these variables, and hence the value for these fields are returned as null. The model will not perform well with missing values, and it is not possible to impute a value for either of these fields when the data items used to calculate them are missing. As a result these two fields are changed to categorical variables, and the value returned for any mortgages with missing values in the calculation will be ‘Unknown’.

5.2.3 Derivation of data items

Much of the research on mortgage arrears and default focuses on negative equity as being one of the most important factors in mortgage arrears and default – especially in the US. It appears that in many cases borrowers who have high levels of negative equity in their home are more likely to default on their mortgage than those who do not have high levels of negative equity. There are a number of ways of calculating both negative equity and current LTV rates for a mortgage. The calculation used as part of this research for the Equity Percent accumulated on the mortgage will be as follows:

$$Equity\ Percentage = \left(\frac{(Current\ Loan\ Balance * -1) - Current\ Property\ Value}{Current\ Property\ Value} \right) * -100$$

This calculation will give the current accumulated equity on the property, returning a minus figure for properties that are in negative equity, and a positive figure for properties where the loan outstanding on the property is less than the value of the property. The Property value is derived by Lender A using the CSO Property Price Index, which is seen as the most useful property price index in Ireland. In some cases the index may undervalue or overvalue properties, but in most cases it is a good estimator of the current value of a particular property.

The current LTV value will be calculated in a very similar way to the Equity Percent value, and both values are very similar. The LTV will be calculated as follows:

$$Current\ Loan\ to\ Value\ (LTV) = \frac{Current\ Loan\ Balance * -1}{Current\ Property\ Value}$$

Whereas the Equity Percent value is a percentage value, the LTV value is returned as a ratio such as 1.25 where the loan balance outstanding is 25% more than what the property is currently worth. The Equity Percent value for this property will be -25% as the loan value is 25% higher than what the property is actually worth. The current LTV and Equity Percentage are correlated, so it is likely that one of the fields will be excluded from the modelling.

It would also be possible to calculate the Equity Percent and LTV values based on the initial loan drawdown value of the mortgages, but it was decided not to use this as the basis for the calculations for a number of reasons. The first reason is that traditionally borrowers would in the past often have released equity in their mortgage to renovate or extend their property, so the initial drawdown amount will not be the full value of the mortgage that the borrower now has. Depending on when the owner drew down the mortgage they may also have paid off a considerable amount of the capital outstanding on the loan, so if current values are taken in to account they may have built up considerable equity in the property. With the recent downward drop in prices, using the drawdown value of the mortgages would also skew the LTV and equity values for a property, as even borrowers who have paid off a considerable portion of the capital would have a very high negative equity percentage based on the purchase price of their home compared to the current value of the property.

5.3 Data Exploration

This section will cover detail some of the structure and content of the data in the dataset and will provide summary statistics for the data. Only a small portion of the data will be described in this section due to space limitations.

5.3.1 Rate of Arrears

There are a total of 70,665 mortgages in the dataset available for building, training and testing the model. Of the total there are 7,525 mortgages in arrears, which equates to approximately 10.5% of the total number of mortgages. The overall mortgage arrears rate in Lender A is higher than this, but many of the mortgages in the bank that are in arrears are Buy to Let (BTL) properties, that are excluded from this research.

In certain areas it is expected that there would be higher than normal arrears rates, as there are certain areas of the country that have been hit with higher unemployment, which can have a bearing on the rate of arrears. The rate of arrears as a percentage of the overall mortgages per county are represented in the Figure 5.1 as calculated and mapped using a Geographical Information System (GIS). This is not necessarily a representative arrears figure for the country as a whole, or indeed the mortgage book of Lender A, but it is indicative of the mortgages in the sample. Figure 5.2 shows the rate of unemployment per county as per the census 2011.

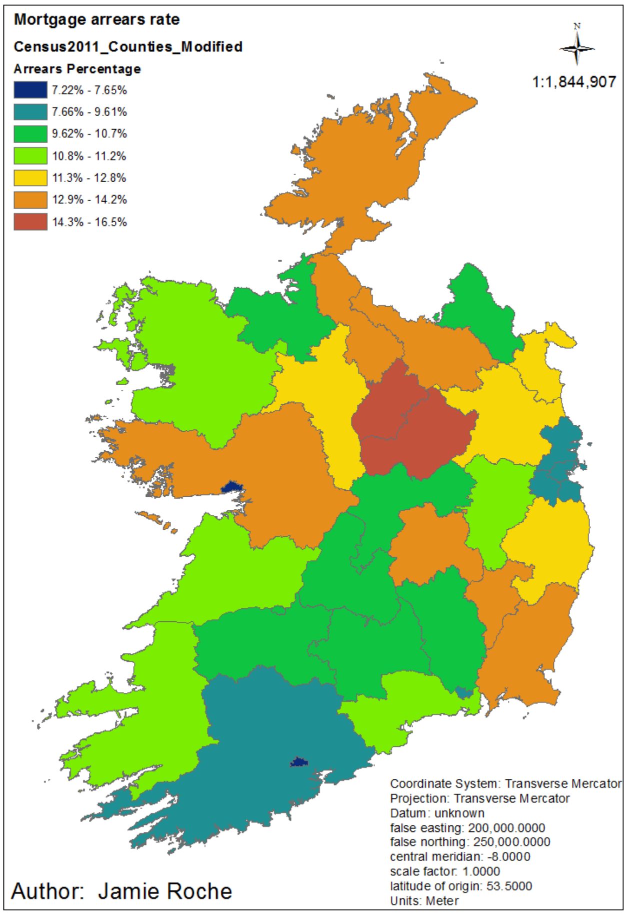


Figure 5.1 Mortgage arrears by county

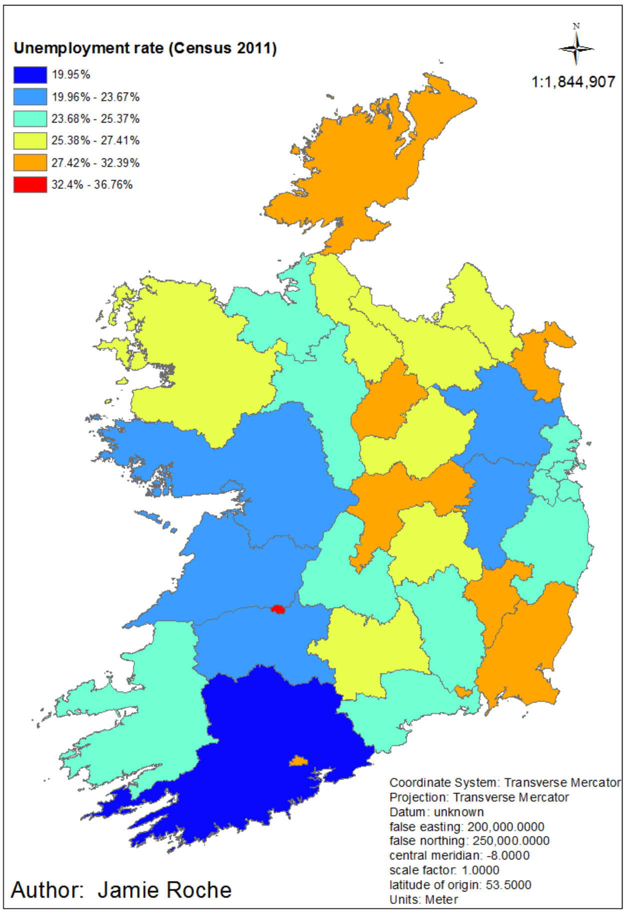

Unemployment rate (Census 2011)
19.95%
19.96% - 23.67%
23.68% - 25.37%
25.38% - 27.41%
27.42% - 32.39%
32.4% - 36.76%
N
1:1,844,907
Coordinate System: Transverse Mercator
Projection: Transverse Mercator
Datum: unknown
false easting: 200,000.0000
false northing: 250,000.0000
central meridian: -8.0000
scale factor: 1.0000
latitude of origin: 53.5000
Units: Meter
Author: Jamie Roche

Figure 5.2 Unemployment rates by county (Census 2011)

For most of the areas with high unemployment from the 2011 census figured, it is apparent that the rate of arrears is significantly higher. Conversely, in most cases where the rate of unemployment is low, the rate of mortgage arrears is also lower.

5.3.2 Spend/Save/Live/Mortgage Transactional Trends

One of the main goals for this experiment is to investigate the predictive capability of the derived spending habits of a customer in an arrears prediction model. The following section details the difference in spending habits for borrowers who are in arrears and those not in arrears, though this may not translate directly in to the models.

Average spend per category for arrears/non-arrears

Figures 5.3 and 5.4 document the average amount per mortgage in each of the Spend/Save/Live/Mortgage categories, with Figure 5.3 displaying the average spend for mortgages that are not in arrears, and Figure 5.4 detailing mortgages that are in arrears.

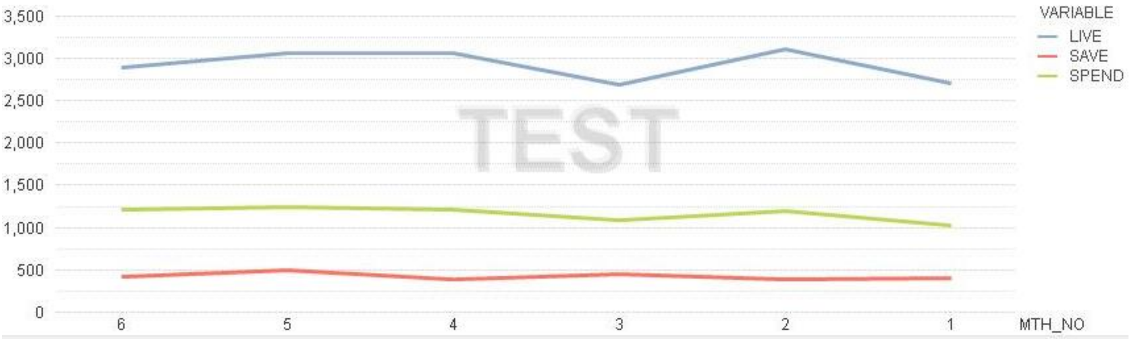


Figure 5.3 Spend/Save/Live breakdown for mortgages not in arrears

As is expected there are differences in the spending habits when looking at the averages across all of the mortgages. The average amount spent in each category is generally higher for the mortgages not in arrears than it is for the mortgages that are in arrears. The average Live expenditure for non-arrears mortgages is between €2,700 and €3,150, whereas for mortgage in arrears the amount is between €2,650 and €3,000 with a noticeable dip in the amount for months 3 and 4 prior to arrears. Even more noticeably different is the change in the Spend category. The average expenditure for non-arrears mortgages of between €1,000 and €1,250 is much greater than the arrears mortgages of between €750 and €1,000. For mortgages in arrears there is an obvious dip in the Spend category three months out from arrears, which would signify the borrower tightening their discretionary spending as they move towards arrears.

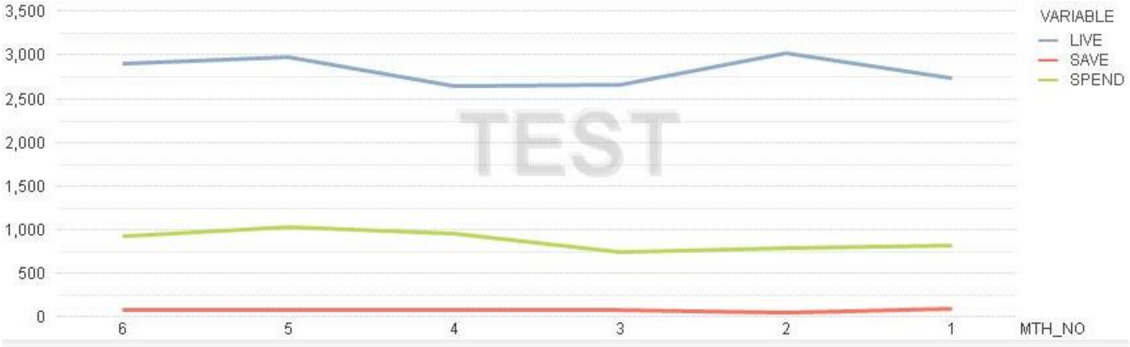


Figure 5.4 Spend/Save/Live breakdown for mortgages in arrears

However, the most striking difference in the spending habits of borrowers in arrears versus borrowers who are not in arrears is in the level of saving. Borrowers who are not in arrears tend to save a consistent amount of approximately €500 per month, but the average savings per month of borrowers who are in arrears is close to zero, even six months in advance of them going in to arrears. This would point to a number of potentially interesting details, such as borrowers in arrears not having any capability to save even before going in to arrears, and when they do get in to financial difficulty they have no savings with which to extricate themselves from the arrears. The total spending across all categories for borrowers in arrears is approximately €3,500 per month, with the borrowers not in arrears spending coming close to €4,750. This would back up the theory that many of the borrowers in difficulty are borrowers who have suffered income shocks, and simply find it more difficult to make ends meet.

Figure 5.5 details the average monthly spending across all categories for the six month period for mortgages in arrears and those not in arrears. The differences between the mortgages in arrears and those not in arrears are quite apparent, with a slight increase in Live, but a very noticeable difference in both Spend and Save categories.

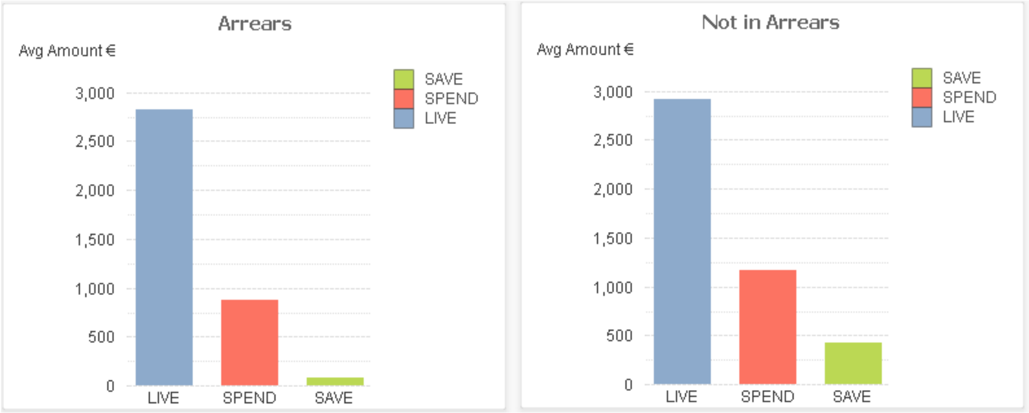


Figure 5.5 Average Spend/Save/Live per month

5.4 *Model Building and Training*

5.4.1 Variable Selection

As there 120 variables in the full dataset created to build the model, it is necessary to reduce the number of variables that are actually used to create the model. The number of training examples required to train the model increases with the number of variables used, which is commonly referred to as the “Curse of Dimensionality”. It stipulates that for every new variable added to a dataset for building a model, more training examples are required to ensure there are sufficient examples of each combination of the variables. If there are 4 categorical variables in a dataset that each have 2 possible outcomes, then there is a total of 16 possible different combinations of these values as the number of combination is 2^d, which in this case is 2^4. If there were 20 variables with two possible outcomes then the number of combinations would be 2^{20} which is just over 1 million combinations. In most datasets though the categorical variables tend to be discrete with a number of different possible outcomes, rather than just binary outcomes, so the number of possible combinations can increase greatly. This is even without taking in to account continuous variables, which can have a near exponential number of different values.

5.4.2 Correlation Matrices

Correlation matrices for the variables within the dataset should reveal any correlations or covariance among the variables. In each of the following matrices variables that are not correlated are shown as having white circles with blue signifying positive correlations, and red negative correlations. If two fields are highly correlated then the representation will be closer to a straight line than a circle.

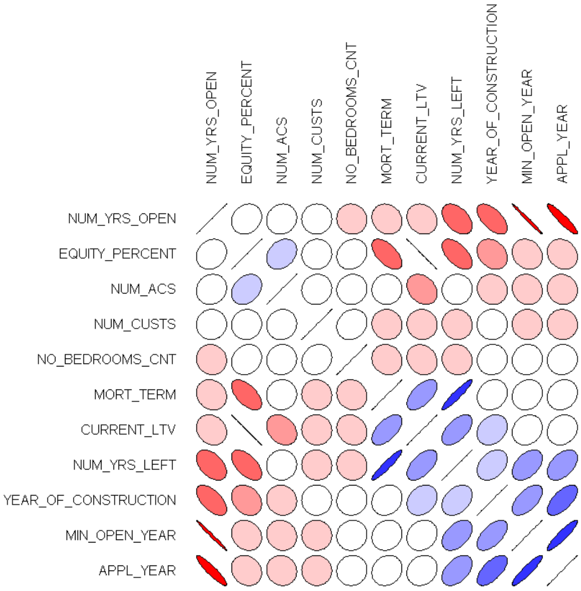


Figure 5.6 Correlation Matrix categorical variables

Figure 5.6 shows the correlation between many of the categorical variables. There is a high level of correlation among many of the variables, but this is as expected as there are many date/year fields included in the dataset that are different variations of the same data. The Num_Yrs_Open and Min_Open_Year are highly negatively correlated as evidence by the near straight line. There is also a very strong negative correlation between the variables Equity Percent and Current_LTV, but again this is expected as the two values are the inverse of one another. Min_Open_Year and Appl_Year have a high degree of positive correlation as evidenced by the shape of the correlation.

Figure 5.7 shows the correlation between some of the continuous variables with an obvious negative correlation existing between the variables Mth_6_Live_Percent and Mth_6_Spend_Percent. A relatively strong negative correlation exists between the Mth_6_Save_Percent variable and the Mth_6_Live_Percent variable, but it is not overly strong. There is also a strong positive correlation between the variables Mth_6_Overlimit_Percent and Mth_6_Utilised_Precent, which would seem appropriate as it would be necessary to have a high utilisation percentage to also have an over limit percentage on a credit card. There are other weak negative correlations, but none that are as obvious or strong enough to conclude that the variables could be removed.

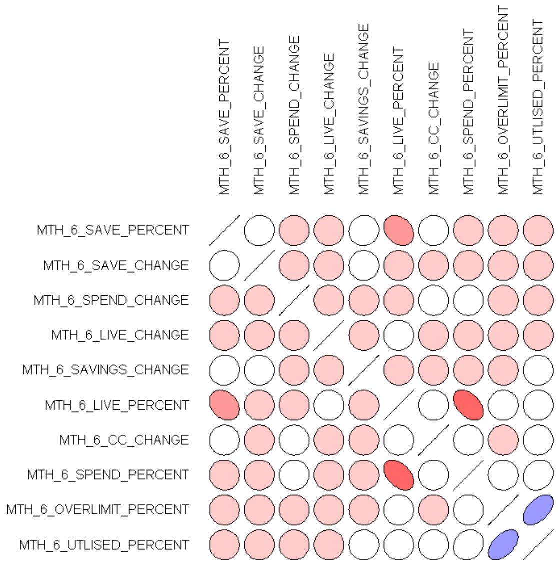


Figure 5.7 Correlation Matrix continuous variables

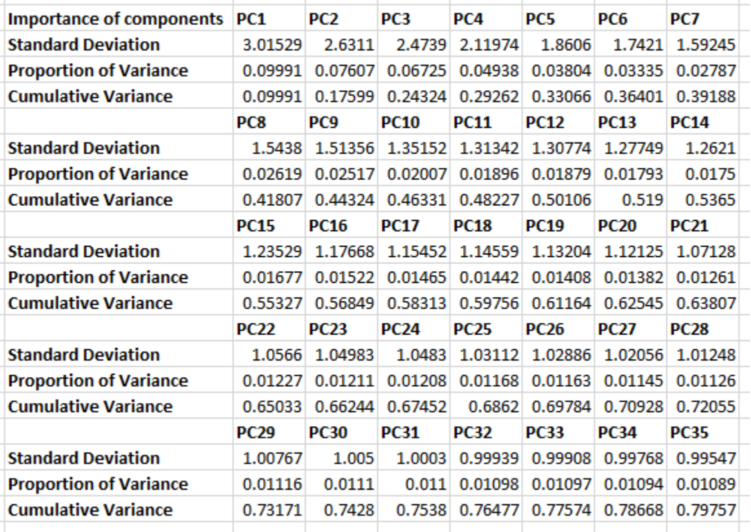
5.4.3 Principal Component Analysis

Principal Component Analysis (PCA) is applied to the dataset to try and reduce the number of variables in the dataset without losing any of the integrity in the data. There are 91 numeric variables in the dataset, and these 91 variables are used to compute the principal components for the dataset. The PCA function was used within Rattle to compute the principal components, and there were 90 principal components identified from the process. Each of the 91 numeric variables in the dataset were used to help generate these principal components, with each of the variables being multiplied by a specific value, with all values added together to create the overall principal component.

As the goal of using PCA is to reduce the number of variables, having ninety principal components does not do much to reduce the number of variables, especially when there are ninety one separate calculations that are make up each one of the components. The PCA function gives a reading of the both the standard deviation and variance for each principal component. These readings are the standard deviation and variance achieved by the component with respect to the whole dataset.

In order to be able to reduce the number of principal components that will be required in the modelling while still keeping the complexities and variance within the dataset, both the standard deviation and variance can be used to identify which components should be used in the dataset. The PCA function lists the components in order of importance. In this case component one is the most important and component ninety the least important. To pick which components should be included in the dataset all components with a standard deviation of 1 or over should be included, or all components with a cumulative variance of above 80%. For the output of the PCA run as part of this experiment the first thirty five components all have a standard deviation of roughly 1 or above, and the combined cumulative variance is 0.79757, so these thirty five components have been chosen, with the other fifty five components discarded. Table 5.1 shows the summary values for the chosen components. By substituting the thirty five principal components for the ninety one numeric variables in the dataset, the overall size of the dataset has been reduced drastically.

Table 5.1 Principal Component Analysis summary

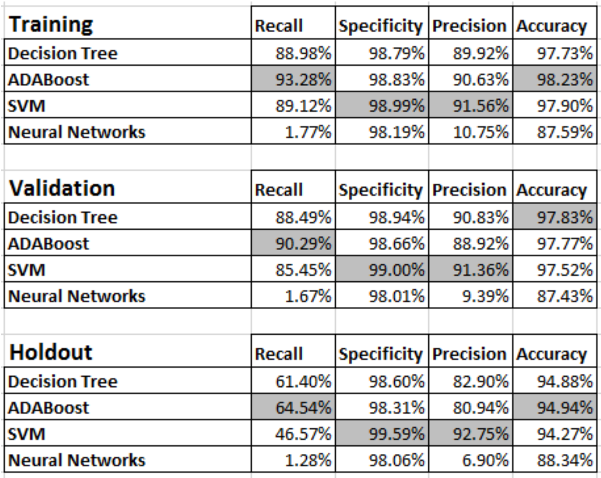


Importance of components	PC1	PC2	PC3	PC4	PC5	PC6	PC7
Standard Deviation	3.01529	2.6311	2.4739	2.11974	1.8606	1.7421	1.59245
Proportion of Variance	0.09991	0.07607	0.06725	0.04938	0.03804	0.03335	0.02787
Cumulative Variance	0.09991	0.17599	0.24324	0.29262	0.33066	0.36401	0.39188
	PC8	PC9	PC10	PC11	PC12	PC13	PC14
Standard Deviation	1.5438	1.51356	1.35152	1.31342	1.30774	1.27749	1.2621
Proportion of Variance	0.02619	0.02517	0.02007	0.01896	0.01879	0.01793	0.0175
Cumulative Variance	0.41807	0.44324	0.46331	0.48227	0.50106	0.519	0.5365
	PC15	PC16	PC17	PC18	PC19	PC20	PC21
Standard Deviation	1.23529	1.17668	1.15452	1.14559	1.13204	1.12125	1.07128
Proportion of Variance	0.01677	0.01522	0.01465	0.01442	0.01408	0.01382	0.01261
Cumulative Variance	0.55327	0.56849	0.58313	0.59756	0.61164	0.62545	0.63807
	PC22	PC23	PC24	PC25	PC26	PC27	PC28
Standard Deviation	1.0566	1.04983	1.0483	1.03112	1.02886	1.02056	1.01248
Proportion of Variance	0.01227	0.01211	0.01208	0.01168	0.01163	0.01145	0.01126
Cumulative Variance	0.65033	0.66244	0.67452	0.6862	0.69784	0.70928	0.72055
	PC29	PC30	PC31	PC32	PC33	PC34	PC35
Standard Deviation	1.00767	1.005	1.0003	0.99939	0.99908	0.99768	0.99547
Proportion of Variance	0.01116	0.0111	0.011	0.01098	0.01097	0.01094	0.01089
Cumulative Variance	0.73171	0.7428	0.7538	0.76477	0.77574	0.78668	0.79757

5.5 Addressing the imbalance in the dataset

To address the imbalance in the dataset between the majority and minority classes, it is necessary to apply a number of methods that should help to address the issue. Each of the methods is discussed briefly below. While the models being built as part of this experiment should benefit from being trained and validated against a balanced dataset, any model that would be produced will always be run against a real-life dataset that is imbalanced in favour of mortgages with no arrears. For this reason the action taken to balance the dataset for training and validation will not be applied in any way to the holdout dataset. Each method will be applied to the training and validation dataset, to work out which method gives the best result and tested against the holdout dataset. It is expected that some of the techniques may work well for some of the algorithms, but will not work well for all of the different algorithms.

Table 5.2 Initial model results



Training	Recall	Specificity	Precision	Accuracy
Decision Tree	88.98%	98.79%	89.92%	97.73%
ADABoost	93.28%	98.83%	90.63%	98.23%
SVM	89.12%	98.99%	91.56%	97.90%
Neural Networks	1.77%	98.19%	10.75%	87.59%

Validation	Recall	Specificity	Precision	Accuracy
Decision Tree	88.49%	98.94%	90.83%	97.83%
ADABoost	90.29%	98.66%	88.92%	97.77%
SVM	85.45%	99.00%	91.36%	97.52%
Neural Networks	1.67%	98.01%	9.39%	87.43%

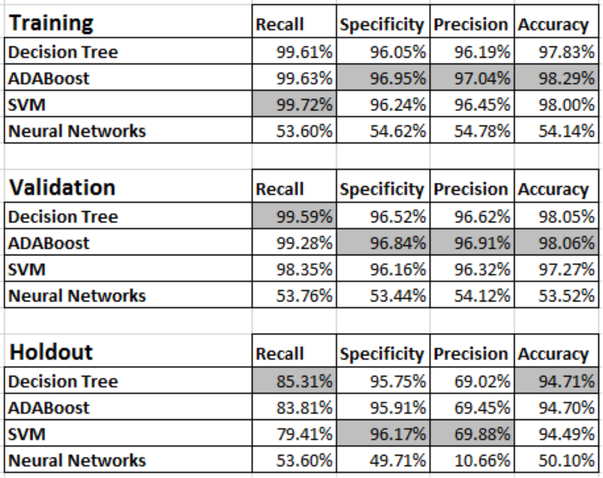
Holdout	Recall	Specificity	Precision	Accuracy
Decision Tree	61.40%	98.60%	82.90%	94.88%
ADABoost	64.54%	98.31%	80.94%	94.94%
SVM	46.57%	99.59%	92.75%	94.27%
Neural Networks	1.28%	98.06%	6.90%	88.34%

Table 5.2 shows the results achieved for the initial modelling of the arrears dataset with all one hundred and twenty variables included in the model building. For each of the measurements (Recall, Specificity, Precision and Accuracy) the best score out of all of the models is highlighted in grey.

5.5.1 Random Undersampling

Random Undersampling is carried out by randomly sampling the available mortgages with no arrears, so that the number of mortgages with arrears matches the number with no arrears. There are 6,538 mortgages in the training and validation dataset in arrears, so a random sample of 6,538 mortgages are taken from the available mortgages with no arrears, and the model is built with the two classes evenly balanced. A number of iterations of random undersampling are undertaken, by sampling without replacing the values, and an average is taken for each of the models built against this balanced dataset.

Table 5.3 Initial results from Random Undersampling



Training	Recall	Specificity	Precision	Accuracy
Decision Tree	99.61%	96.05%	96.19%	97.83%
ADABoost	99.63%	96.95%	97.04%	98.29%
SVM	99.72%	96.24%	96.45%	98.00%
Neural Networks	53.60%	54.62%	54.78%	54.14%

Validation	Recall	Specificity	Precision	Accuracy
Decision Tree	99.59%	96.52%	96.62%	98.05%
ADABoost	99.28%	96.84%	96.91%	98.06%
SVM	98.35%	96.16%	96.32%	97.27%
Neural Networks	53.76%	53.44%	54.12%	53.52%

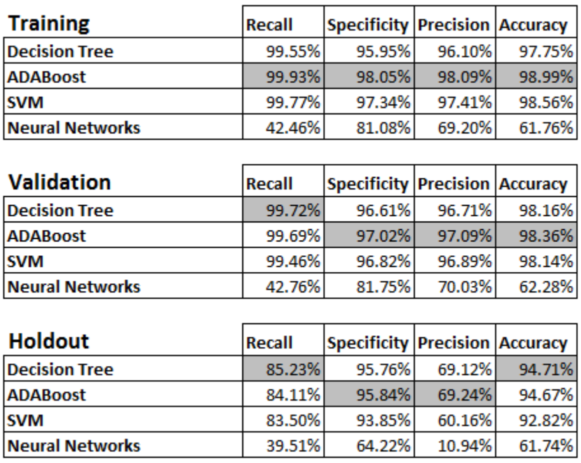
Holdout	Recall	Specificity	Precision	Accuracy
Decision Tree	85.31%	95.75%	69.02%	94.71%
ADABoost	83.81%	95.91%	69.45%	94.70%
SVM	79.41%	96.17%	69.88%	94.49%
Neural Networks	53.60%	49.71%	10.66%	50.10%

Table 5.3 details the initial results from three iterations of the Random Undersampling, with the averages taken for each of the models produced. There is a marked increase in the Recall values for the models on the holdout data, but the Precision values are still quite low, with the number of False Positives impacting greatly on this figure. The Random Undersampling has allowed the model to gain a greater understanding of the minority class, without sacrificing how it performs on the majority class predictions, as the Specificity rate has only dropped about approximately 2-3% on the holdout data. The recall score for the decision tree of 85.31% on the holdout dataset is a very acceptable score for a prediction model, but this may be improved upon by other processes.

5.5.2 Random Oversampling

To test out whether or not Random Oversampling can help to produce a more accurate model, the minority class of mortgages with arrears will be oversampled in the training and validation dataset. This will be achieved by duplicating entries in the dataset for instances where a mortgage has arrears. In order to totally address the balance of mortgages in arrears with those not in arrears with Random Undersampling alone, each mortgage in arrears would need to be included in the dataset approximately ten times. By including each mortgage ten times in the sample the model will likely be overfitted to the training data and perform badly against the holdout data. Random Oversampling by itself may not produce a beneficial result, so oversampling will also be used in conjunction with undersampling, so that a smaller number of records will need to be generated from the oversampling. Random oversampling will be tested in using three different samples: the first where the number of minority class items is doubled; the second where the number of minority class items is quadrupled; and the third where the number of minority class items is doubled with undersampling being used so that the minority class has double the number of records than the majority class.

Table 5.4 Initial results from oversampling



Training	Recall	Specificity	Precision	Accuracy
Decision Tree	99.55%	95.95%	96.10%	97.75%
ADABoost	99.93%	98.05%	98.09%	98.99%
SVM	99.77%	97.34%	97.41%	98.56%
Neural Networks	42.46%	81.08%	69.20%	61.76%

Validation	Recall	Specificity	Precision	Accuracy
Decision Tree	99.72%	96.61%	96.71%	98.16%
ADABoost	99.69%	97.02%	97.09%	98.36%
SVM	99.46%	96.82%	96.89%	98.14%
Neural Networks	42.76%	81.75%	70.03%	62.28%

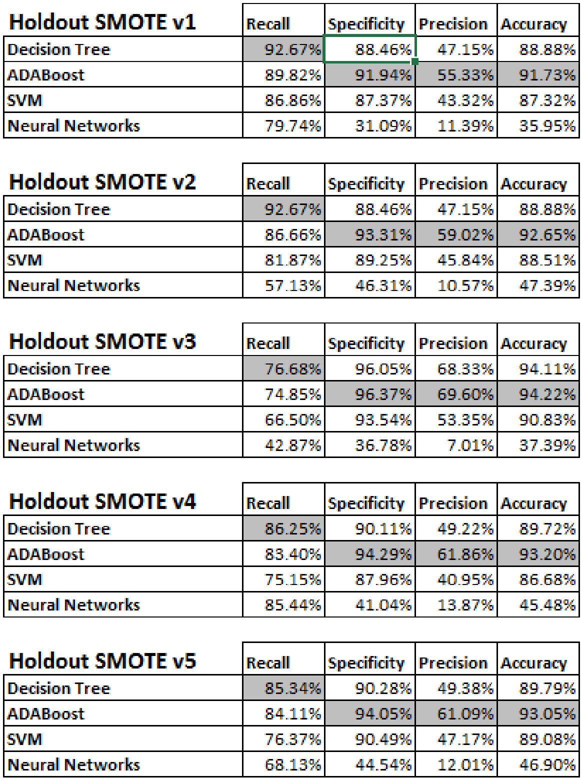
Holdout	Recall	Specificity	Precision	Accuracy
Decision Tree	85.23%	95.76%	69.12%	94.71%
ADABoost	84.11%	95.84%	69.24%	94.67%
SVM	83.50%	93.85%	60.16%	92.82%
Neural Networks	39.51%	64.22%	10.94%	61.74%

Table 5.4 details the results best results achieved using oversampling to address the class imbalance. The minority class has been oversampled by 100%, so that the number of arrears mortgages in the training dataset is 13,076 and the number of mortgages not in arrears has been undersampled, so that there are also 13,076 in the dataset. Most of the models with the exception of the Neural Networks perform very well on the training and validation data and also perform well on the holdout data. The AdaBoost algorithm performs best in most of the iterations, though the score of the decision tree are very closely similar. Ultimately, the decision tree model performs best on the recall for the holdout dataset, so it is likely that this would be the model chosen if oversampling is chosen as the best sampling technique to address the imbalance in the dataset.

5.5.3 Synthetic Sampling

Synthetic Sampling will be used to produce new data that will supplement the existing minority class data. The SMOTE algorithm will be used in R to produce synthetic data that is similar, but not identical to the existing data items. To test that Synthetic samples improve the model a number of iterations using synthetic data will be completed, with a varying range of synthetic samples created. The number of synthetic examples created will be in the range from the number of current minority class samples right up to the total number of examples required to balance the mortgages with arrears with those with no arrears (6,538 up to 54,241). It is expected that the models trained using SMOTE should perform well, though it may be at the risk of over-fitting the training and validation data.

Table 5.5 SMOTE Synthetic sampling holdout scores



Holdout SMOTE v1	Recall	Specificity	Precision	Accuracy
Decision Tree	92.67%	88.46%	47.15%	88.88%
ADABoost	89.82%	91.94%	55.33%	91.73%
SVM	86.86%	87.37%	43.32%	87.32%
Neural Networks	79.74%	31.09%	11.39%	35.95%

Holdout SMOTE v2	Recall	Specificity	Precision	Accuracy
Decision Tree	92.67%	88.46%	47.15%	88.88%
ADABoost	86.66%	93.31%	59.02%	92.65%
SVM	81.87%	89.25%	45.84%	88.51%
Neural Networks	57.13%	46.31%	10.57%	47.39%

Holdout SMOTE v3	Recall	Specificity	Precision	Accuracy
Decision Tree	76.68%	96.05%	68.33%	94.11%
ADABoost	74.85%	96.37%	69.60%	94.22%
SVM	66.50%	93.54%	53.35%	90.83%
Neural Networks	42.87%	36.78%	7.01%	37.39%

Holdout SMOTE v4	Recall	Specificity	Precision	Accuracy
Decision Tree	86.25%	90.11%	49.22%	89.72%
ADABoost	83.40%	94.29%	61.86%	93.20%
SVM	75.15%	87.96%	40.95%	86.68%
Neural Networks	85.44%	41.04%	13.87%	45.48%

Holdout SMOTE v5	Recall	Specificity	Precision	Accuracy
Decision Tree	85.34%	90.28%	49.38%	89.79%
ADABoost	84.11%	94.05%	61.09%	93.05%
SVM	76.37%	90.49%	47.17%	89.08%
Neural Networks	68.13%	44.54%	12.01%	46.90%

The synthetic sampling carried out to address the class imbalance was implemented using the Synthetic Minority Over-sampling Technique (SMOTE) technique, available in the R package DMwR[3]. Synthetic data was created for the minority class and then combined with the majority class data that would be used to generate the model. In all cases the prediction of the target class improved using the synthetic data, but the prediction of the majority class suffered to a large degree. All of the models reported a much higher recall score, as it was easier for the models to identify the arrears cases,

[3] http://cran.r-project.org/web/packages/DMwR/index.html

with the Decision Tree model reporting a score of 92.67%. However, all of the models also reported a significant decline in the Precision score, which can be directly attributed to the number of false positives from the mortgages with no arrears. It would appear that the synthetic data created has allowed the greater identification of the mortgages with arrears, but at the expense of being able to correctly identify the non-arrears mortgages.

5.6 Misclassification Costs

When building a model such as this arrears prediction model, it is common to apply a cost to the misclassification of examples in the dataset. These misclassification costs can be used in the building of the models to work out which model has the best performance, and it can be used to differentiate between models that have similar overall prediction scores, but perform differently on the constituent classes in the data.

Normally a higher cost is associated with the misclassification of positive classes (False Negatives), and in this experiment a much higher cost will be associated with mortgages that are in arrears, but have been classified as mortgages that do not have arrears. For mortgages that are not in arrears but have been misclassified as being in arrears (False positives), the associated misclassification cost will be much lower. No cost will be associated with mortgages that are classified correctly, as the Arrears Support Unit (ASU) will not need to engage with these borrowers.

The higher costs associated with the False Negatives is due to the increased cost of managing the arrears on a mortgage if they were not foreseen, as well as any actual losses that may be realised by the Lender. For the purpose of this experiment the costs associated with misclassifying the outcome for an individual mortgage will be as follows:

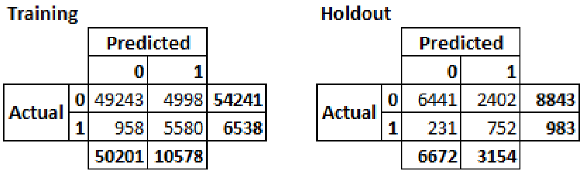
False Positives - €80
False Negatives - €230

These costs have been reached in consultation with the Arrears Support Unit, by quantifying the cost associated with engaging with borrowers who are in difficulty. The much lower cost associated with the False Positives covers doing some background research on the customer as well as contacting the customer. The cost for the False Negatives covers having to contact the customer a number of times as a result of not capturing the arrears before it has happened, whereby the customer will already have gone in to arrears before the bank has contacted them. Neither of these two misclassification costs take in to account potential future losses for the Lender from arrears, as there are many different solutions being put in place for borrowers that can result in varying costs or losses for the bank.

The misclassification costs will be applied to the error matrices produced by the models built on the data, and will be used as another measure of how well each model performs. Models that perform better on the minority class in the data should give the lowest overall misclassification costs but the False Positive rate will also be a factor in the costs.

5.7 Naive Bayes Classifier

All of the same data available for the building and training of the models in R is made available to build the classifier in Aster using Naïve Bayes. Each of the records in the combined training/validation dataset is used to build and test the model, and then the model is applied to the holdout dataset to assess the predictive capability of the model. Figure 5.8 shows the confusion matrix for the output, while Figure 5.9 gives the output for each of the formulas that are being used to evaluate all of the other models created.



Training

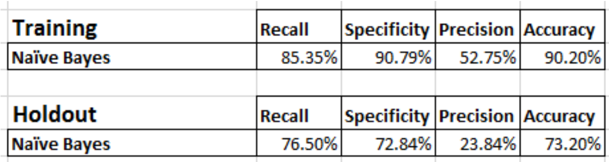
		Predicted		
		0	1	
Actual	0	49243	4998	54241
	1	958	5580	6538
		50201	10578	

Holdout

		Predicted		
		0	1	
Actual	0	6441	2402	8843
	1	231	752	983
		6672	3154	

Figure 5.8 Confusion Matrix for Naïve Bayes classifier

The recall rate for the Naïve Bayes classifier is relatively high for the model training, but this drops by quite a large amount when testing against the holdout data sample. The precision of the model is quite low at 52.75% for the training data and just 23.84% for the holdout dataset, due in large part to the number of false positives generated, along with the relatively high number of false negatives. It’s apparent that the Naïve Bayes classifier is not suitable for this classification problem. With the training scores as low as they are, it is unlikely that the model could be improved to such as level that would bring the misclassifications to within acceptable tolerances. The reason for this is likely to be that a Naïve Bayes classifier assumes independence among the variables, whereas many of the variables in the dataset are very closely dependent, especially the Spend/Save/Live/Mortgage spending categories which are highly correlated, as each category is a percentage of the total spending per month.



Training	Recall	Specificity	Precision	Accuracy
Naïve Bayes	85.35%	90.79%	52.75%	90.20%

Holdout	Recall	Specificity	Precision	Accuracy
Naïve Bayes	76.50%	72.84%	23.84%	73.20%

Figure 5.9 Naïve Bayes classifier evaluation scores

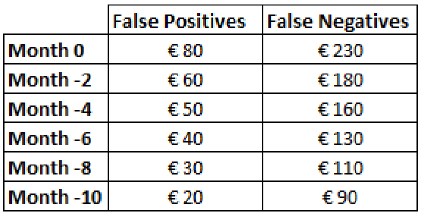
5.8 Finding the best time to predict arrears

To properly engage with borrowers in difficulty to limit the extent of arrears it is necessary to know in advance if a borrower is in distress or not. Once the best predictive model is produced for the holdout dataset, models will be produced that will try to predict a number of months in advance what mortgages are likely to go in to arrears. In a similar way to the standard model being applied to the data, all of the data items will be extracted from the data warehouse and fed in to the model, with the difference being that the data will be extracted from a number of months in advance of the arrears event.

If there is sufficient time, the model will be run for each of the calendar months in the year prior to the arrears event, but if this is not feasible then it will be completed against every second month in the year. The actual model used for predicting arrears in advance should be very similar to the model used to predict arrears one month in advance, though the model may place a higher importance on different data items.

The known outcome of each of the mortgages will again be used to train the model, and the probability given for each prediction along with the predicted class will be used to calculate how accurate the model is. It is likely that the characteristics of a mortgage twelve months in advance of arrears is very different to that of a mortgage when it is only one or two months in advance of arrears, but it is difficult to quantify this as each borrower and mortgage is different. Consequently it may be apparent that the model’s predictive capabilities may be diminished when looking the mortgages a number of months in advance of arrears. A revised misclassification cost matrix will be applied to the model that will be tasked with predicting arrears a number of months in advance of arrears, to reflect the additional time element provided by predicting the event in advance. This revised cost matrix will be used to aid in the decision process when comparing the outputs of the multiple models to find the optimum time for predicting arrears in advance. The revised misclassification costs are as follows:

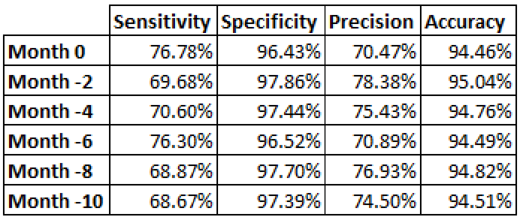
Table 5.6 Misclassification costs for advance predictive models



	False Positives	False Negatives
Month 0	€ 80	€ 230
Month -2	€ 60	€ 180
Month -4	€ 50	€ 160
Month -6	€ 40	€ 130
Month -8	€ 30	€ 110
Month -10	€ 20	€ 90

Table 5.6 details the misclassification costs associated with each of the models that are going to be run to try to predict arrears in advance of the event. The costs associated with misclassification closer to the actual arrears event would be closer to the original costs of €80 per False Positive and €230 per False Negative, but for those misclassified six or ten months in advance of the arrears, the cost would be much lower. The costs of the lender contacting the borrower a number of months in advance of them potentially going in to arrears are lower as a method such as a letter could be used instead of a phone call, which would be required when the arrears are potentially imminent. However, the same exposure is there is the model fails to correctly classify a mortgage that is going to go in to arrears, when the event could potentially be avoided. While the costs associated with these models are much more hypothetical than the costs associated with the regular model, they still provide a way of measuring the benefit of the models.

Table 5.7 Predicting arrears in advance evaluation scores

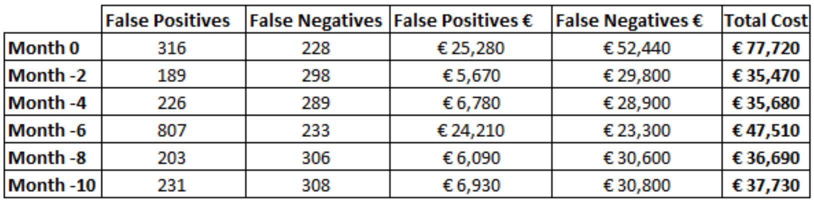


	Sensitivity	Specificity	Precision	Accuracy
Month 0	76.78%	96.43%	70.47%	94.46%
Month -2	69.68%	97.86%	78.38%	95.04%
Month -4	70.60%	97.44%	75.43%	94.76%
Month -6	76.30%	96.52%	70.89%	94.49%
Month -8	68.87%	97.70%	76.93%	94.82%
Month -10	68.67%	97.39%	74.50%	94.51%

Table 5.7 shows the results achieved when predicting arrears a number of months in advance of the actual arrears event. Month 0 is the current month that is used for all of the models already completed and discussed earlier in the chapter. Each of the additional months included in the table is the model being created for the same subset of mortgages,

where the data used is the data taken *n* months in advance of arrears, where *n* is the number of month in question. For mortgages not in arrears the data is taken *n* months prior to point that the data was extracted (i.e. September 2013 for the training data and February 2014 for the holdout data). The models have been built and trained at a bi-monthly interval to test when the most appropriate time is to predict the arrears, to allow the lender to intervene and potentially stop the borrower going in to arrears in the first place.

Table 5.8 Predicting arrears in advance - numbers and misclassification costs



	False Positives	False Negatives	False Positives €	False Negatives €	Total Cost
Month 0	316	228	€ 25,280	€ 52,440	**€ 77,720**
Month -2	189	298	€ 5,670	€ 29,800	**€ 35,470**
Month -4	226	289	€ 6,780	€ 28,900	**€ 35,680**
Month -6	807	233	€ 24,210	€ 23,300	**€ 47,510**
Month -8	203	306	€ 6,090	€ 30,600	**€ 36,690**
Month -10	231	308	€ 6,930	€ 30,800	**€ 37,730**

5.9 Conclusions

This chapter detailed the experiment carried out to build a prediction model capable of predicting mortgage arrears before they happen. For the model to have a high degree of accuracy a number of different models were built with different algorithms and with different sampling techniques. The different sampling techniques were necessary to address the class imbalance in the data, as there is a very high ratio of mortgages in the dataset that have no arrears. The goal of using different sampling techniques to address the imbalance in the dataset are to improve the classification of the mortgages in arrears, without compromising on the prediction of the non-arrears mortgages. The techniques and results used for addressing the imbalance will be discussed further in chapter six.

The chapter also presented the misclassification costs that will be applied to each of the models, as well as some of the other measures that will be used to evaluate the models created. Chapter six will evaluate these models in more detail whereby the overall costs of each model can be used to help choose which model is the most appropriate, along with the other measures such as Recall and Precision.

The following chapter will discuss in more detail the results achieved during the experiment, and will evaluate each of the sampling and modelling techniques in more detail.

6. EVALUATION

6.1 Introduction

This chapter will focus on evaluating the results achieved from the experiment, as well as discussing the overall research carried out as part of the thesis. The various techniques applied to the data and modelling phases to achieve a better predictive model will be evaluated and ranked based on the outcomes achieved.

The results achieved from the modelling experiment will be assessed and interpreted based on the research carried out in the literature review sections previously discussed.

One of the goals of the experiment is to investigate if the addition of newly available transaction level data long with data on negative equity, can help to increase the predictive capability of the models.

6.2 Results Evaluation

6.2.1 Evaluating the models produced in the experiment

All of the models produced as part of this experiment will be evaluated on the holdout dataset using a number of different measurements, which will gauge how affective the models have been on different aspects of the classification. The scores will be calculated for the Recall, Specificity, Precision and Accuracy of each of the models. Each modelling technique employed as part of the research experiment will work in different ways, so it is necessary to be able to compare all of the results with measurements that will consistently be able to compare and contrast the performance of the individual models. Each of the four measurements mentioned earlier is calculated as follows:

$$Recall = \frac{TP}{TP + FN} \qquad Specificity = \frac{TN}{FP + TN}$$

$$Precision = \frac{TP}{TP + FP} \qquad Accuracy = \frac{TP + TN}{TP + FP + TN + FN}$$

The most important of the evaluation methods for the models will be the Recall score. This score will determine how each model has performed at predicting the class of interest, which are the mortgages that are in arrears. The recall score is the percentage of the target class that have been correctly classified, and details how well the model has performed in predicting the target class, which in this case is very much the minority class. The Specificity rate defines how well the negative class has been predicted, while the Precision rate details how accurate the positive predictions of the model are.

The Accuracy of the model in this case is the overall predictive accuracy, which is the sum of all accurate predictions over the total number of records in the dataset. The Accuracy of the model will be the least used of the 4 measures, as the imbalance in the dataset can cause the accuracy to be high, without the target class performing well. It is expected that there will be quite a large variance in how the different models perform under different conditions, especially using the different sampling methods, and when the principal components are taken in to account instead of all of the individual numeric data items.

Table 6.1 details each of the scores for all of the models produced, broken down by each of the different algorithms. For each iteration of the modelling the best Recall, Specificity, Precision and Accuracy score are highlighted in grey across all four different models.

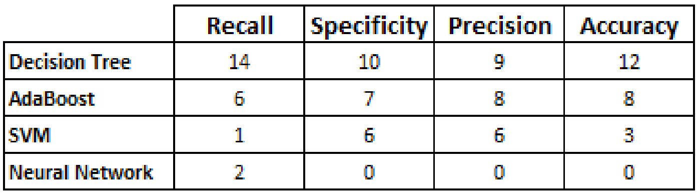
Table 6.1 All model evaluations

	Decision Tree				AdaBoost				SVM				Neural Network			
	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy
Basic Model with no transactional data	0.61	0.99	0.83	0.95	0.66	0.98	0.80	0.95	0.48	1.00	0.92	0.94	0.60	0.98	0.75	0.94
Initial model with all data items included	0.74	0.97	0.73	0.95	0.72	0.97	0.75	0.95	0.60	0.97	0.69	0.93	0.57	0.38	0.09	0.40
Undersampling run 1 (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.85	0.96	0.69	0.95	0.85	0.93	0.57	0.92	0.58	0.34	0.09	0.37
Undersampling run 2 (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.85	0.96	0.69	0.95	0.85	0.93	0.57	0.92	0.86	0.62	0.20	0.65
Undersampling run 3 (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.85	0.96	0.68	0.95	0.85	0.93	0.56	0.92	0.54	0.41	0.09	0.43
Oversampling run 1 (arrears doubled - 13076/6538)	0.85	0.96	0.69	0.95	0.86	0.95	0.68	0.94	0.87	0.93	0.57	0.92	1.00	0.00	0.10	0.10
Oversampling run 2 (arrears tripled - 19614/6538)	0.85	0.96	0.69	0.95	0.87	0.95	0.64	0.94	0.87	0.92	0.55	0.92	0.72	0.39	0.12	0.42
Oversampling run 3 (arrears doubled, equivalent non-arrears - 13076/13076)	0.85	0.96	0.69	0.95	0.84	0.96	0.69	0.95	0.84	0.94	0.60	0.93	0.40	0.64	0.11	0.62
Naïve Bayes all data items	0.77	0.73	0.24	0.73												
Synthetic Sampling 1 (13076 arrears to 6538 non-arrears)	0.93	0.88	0.47	0.89	0.90	0.92	0.55	0.92	0.87	0.87	0.43	0.87	0.80	0.31	0.11	0.36
Synthetic Sampling 2 (13076 arrears to 13076 non-arrears)	0.93	0.88	0.47	0.89	0.87	0.93	0.59	0.93	0.82	0.89	0.46	0.89	0.57	0.46	0.11	0.47
Synthetic Sampling 3 (13076 arrears to 52451 non-arrears)	0.77	0.96	0.68	0.94	0.75	0.96	0.70	0.94	0.66	0.94	0.53	0.91	0.43	0.37	0.07	0.37
Synthetic Sampling 4 (26152 arrears to 26152 non-arrears)	0.86	0.90	0.49	0.90	0.83	0.94	0.62	0.93	0.75	0.88	0.41	0.87	0.85	0.41	0.14	0.45
Synthetic Sampling 5 (26152 arrears to 26152 non-arrears, k=5)	0.85	0.90	0.49	0.90	0.84	0.94	0.61	0.93	0.76	0.90	0.47	0.89	0.68	0.45	0.12	0.47
Initial model with PCA variables added	0.77	0.96	0.71	0.94	0.72	0.97	0.72	0.94	0.50	0.99	0.88	0.94	0.53	0.36	0.08	0.38
Undersampling run 1 PCA (50/50 majority/minority	0.84	0.96	0.69	0.95	0.85	0.96	0.69	0.95	0.83	0.96	0.68	0.94	0.42	0.45	0.08	0.45
Undersampling run 2 PCA (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.86	0.95	0.65	0.94	0.83	0.96	0.68	0.94	0.67	0.40	0.11	0.43
Undersampling run 3 PCA (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.84	0.96	0.69	0.95	0.83	0.96	0.68	0.94	0.61	0.43	0.11	0.45
Oversampling run 1 PCA (number of arrears doubled)	0.85	0.96	0.69	0.95	0.86	0.95	0.68	0.94	0.86	0.95	0.64	0.94	0.60	0.48	0.11	0.49
Oversampling run 2 PCA (arrears quadrupled)	0.85	0.96	0.69	0.95	0.87	0.95	0.65	0.94	0.87	0.93	0.59	0.93	0.87	0.88	0.46	0.88
Oversampling run 3 PCA (2 to 1 arrears to non-arrears) s	0.85	0.96	0.69	0.95	0.84	0.96	0.69	0.95	0.81	0.96	0.70	0.95	0.62	0.20	0.08	0.24
Synthetic Sampling 1 PCA (13076 arrears to 6538 non-arrears)	0.93	0.88	0.47	0.89	0.88	0.92	0.56	0.92	0.84	0.95	0.64	0.94	0.58	0.37	0.09	0.39
Synthetic Sampling 2 PCA (13076 arrears to 13076 non-arrears)	0.93	0.88	0.47	0.89	0.87	0.93	0.59	0.93	0.79	0.95	0.66	0.94	0.58	0.29	0.08	0.32
Synthetic Sampling 3 PCA (13076 arrears to 54541 non-arrears)	0.79	0.93	0.56	0.92	0.75	0.96	0.67	0.94	0.61	0.98	0.74	0.94	0.48	0.34	0.07	0.35

6.2.2 Best Modelling technique for the experiment

The modelling techniques that performed best at predicting what mortgages are going to go in to arrears in most cases were Decision Trees and the AdaBoost algorithm. Decision Trees are often overlooked by many when creating new classification or prediction models, as they are seen as being too simplistic. AdaBoost works by continually focusing on the misclassified examples in the dataset to try to improve the classification of these items. In the different iterations of the model, decision trees have returned the best Recall score for fourteen of the twenty three iterations, and AdaBoost has returned the best score in six of the iterations. In many cases the two scores have been very close for each of the techniques, with only a very small difference between the two.

Table 6.2 Model comparisons by aggregate number of times each model performed best



	Recall	Specificity	Precision	Accuracy
Decision Tree	14	10	9	12
AdaBoost	6	7	8	8
SVM	1	6	6	3
Neural Network	2	0	0	0

The Decision Tree models also perform well on the negative class, with a high score on the Specificity rate in most cases, and it has the highest score for ten of the iterations. The Decision Trees do not seem to work as well at predicting the negative class in the models where synthetic data has been used, as most of these return a sore of less than 90% for the Specificity. The AdaBoost model also performs well on the negative class, with a comparative or better score on the Specificity in most cases. For seven of the iterations, AdaBoost scores higher on Specificity and in general it performs very well on the negative class, with a score of over 90% for all of the model iterations.

The Support Vector Machines performed relatively well in some of the iterations, with a Recall rate that was comparative to the Decision Trees and AdaBoost models, but there were other iterations where the Recall rate was close to or less than 50%. The SVM models did have the highest Specificity and Precision scores in six of the iterations carried out, so while they didn't perform the best on Recall rate in most cases, they did score highly in other areas.

The Neural Networks performed poorly in comparison to the other methods, especially when the data was imbalanced. They did however have the highest Recall score for two of the model iterations, but both were where the data was more balanced than the full dataset. The Precision score in all but a very small number of the models was very low, which would signify that the number of false positives would be extremely high and this would add significantly to the model misclassification costs.

Decision Trees created as part of this experiment have performed to a high level using all of the above measurements, and have been the most consistent performer of all of the models. Together with the easy understanding that comes from the rules behind a Decision Tree, it is likely that Decision Trees would be the best method for predicting mortgages that are likely to go in to arrears. Using the rules created as part of the model building it would be possible to implement these rules in the data warehouse to output the scores, without having to have a separate tool for the prediction.

However, the true costs of the model misclassification rates will be taken in to account when choosing the best model for predicting the arrears, and together with the measurements discussed here, the best model will be chosen.

6.2.3 Optimum sampling approach for addressing the class imbalance

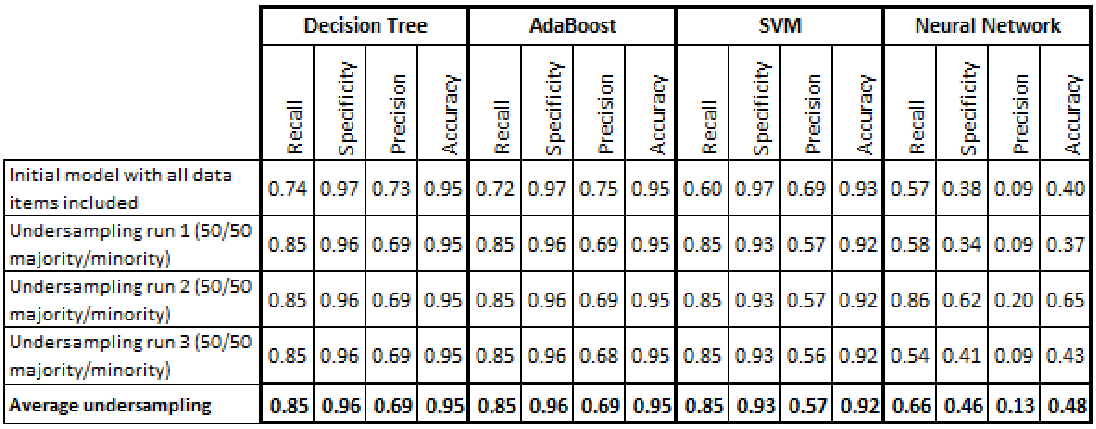
In order to address the imbalance in the two classes in the dataset, a number of sampling approaches have been used to try to build the models with a more balanced dataset. Each of the sampling approaches will be compared and evaluated to investigate if it is possible to achieve a good prediction score with the imbalanced data, by employing a sampling method.

Random Undersampling

Random undersampling was applied to the dataset by taking a random sample of the majority class instead of using all of the records from the dataset. Three iterations of random undersampling were applied to the dataset, each time without replacement, so that three distinct random samples of the majority class were chosen from the data. In each case the training and validation data was made up of the random sample of the majority class along the entire minority class data so that the ratio of minority/majority in the dataset was 50/50.

Table 6.3 details the scores achieved across each of the four measurements for the models created using random undersampling. There are considerable improvements in the Recall rate for most of the models and iterations, though some of the other scores have suffered as a consequence. Both Specificity and Precision has declined for all of the iterations with random undersampling, with the exception of one of the iterations of the Neural Networks. The overall Accuracy score for the models in most cases has not changed very much, but that can be attributed to both the increase in the scores for Recall and the drop in scores for the Specificity and Precision.

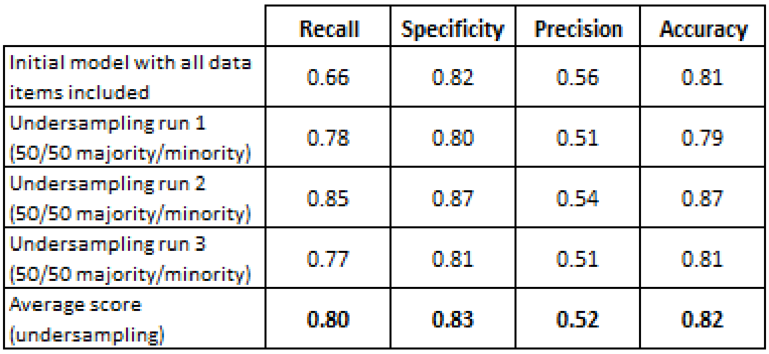
Table 6.3 Undersampling results



	Decision Tree				AdaBoost				SVM				Neural Network			
	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy
Initial model with all data items included	0.74	0.97	0.73	0.95	0.72	0.97	0.75	0.95	0.60	0.97	0.69	0.93	0.57	0.38	0.09	0.40
Undersampling run 1 (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.85	0.96	0.69	0.95	0.85	0.93	0.57	0.92	0.58	0.34	0.09	0.37
Undersampling run 2 (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.85	0.96	0.69	0.95	0.85	0.93	0.57	0.92	0.86	0.62	0.20	0.65
Undersampling run 3 (50/50 majority/minority)	0.85	0.96	0.69	0.95	0.85	0.96	0.68	0.95	0.85	0.93	0.56	0.92	0.54	0.41	0.09	0.43
Average undersampling	**0.85**	**0.96**	**0.69**	**0.95**	**0.85**	**0.96**	**0.69**	**0.95**	**0.85**	**0.93**	**0.57**	**0.92**	**0.66**	**0.46**	**0.13**	**0.48**

Table 6.4 details the average scores across all of the models produced using undersampling. The average Recall rate has jumped considerably to 80% from 66%, with the best score being considerably higher at 85%. The average Specificity score is only marginally higher at 83%, but this is buoyed by one large jump in Specificity from the 2nd iteration of the Neural Networks.

Table 6.4 Average results for undersampling



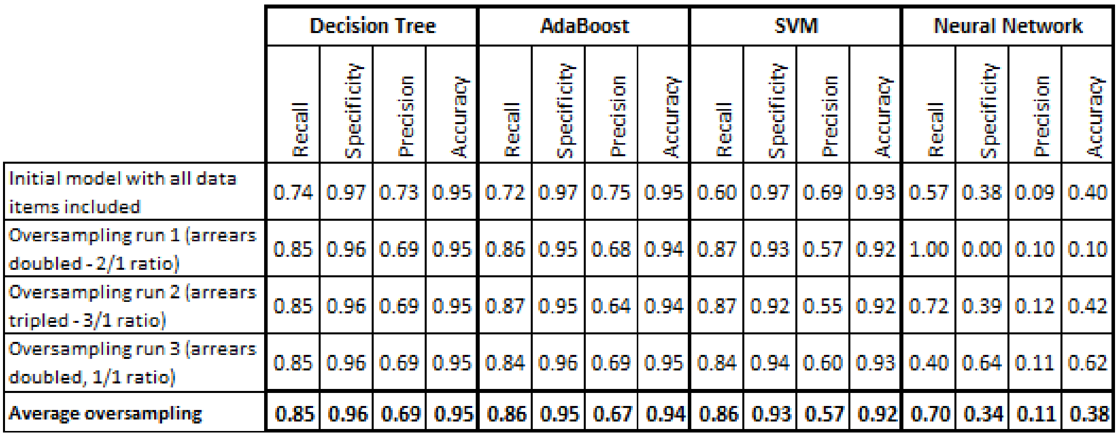
	Recall	Specificity	Precision	Accuracy
Initial model with all data items included	0.66	0.82	0.56	0.81
Undersampling run 1 (50/50 majority/minority)	0.78	0.80	0.51	0.79
Undersampling run 2 (50/50 majority/minority)	0.85	0.87	0.54	0.87
Undersampling run 3 (50/50 majority/minority)	0.77	0.81	0.51	0.81
Average score (undersampling)	**0.80**	**0.83**	**0.52**	**0.82**

An apparent issue with random undersampling is that by taking a random sample of the data it is possible that the sample chosen will differ greatly from the minority class, and hence the models will be able to separate the classes easier. This appears to have happened with the second iteration of the undersampling, as the Neural Networks which had previously performed poorly in almost all of the measurements scored much higher. Overall though the random sampling method has produced much higher scores, at the risk of losing some accuracy in the Specificity and Precision rates for the majority class.

Random Oversampling

The oversampling of the minority class data was carried out by duplicating records from the minority class in the dataset, in order to increase the number of minority class records available for training and validation. The sampling was not random per se, as all records in the dataset were duplicated in order to increase the number of records in the dataset. It was felt that by duplicating all records in the minority class by a certain factor would produce a better result than just selecting a random sample of the records. This was in part due to the overwhelming difference between the total minority class and majority class records, where a simple sample of the minority class records would be insufficient to balance the dataset. Three iterations of oversampling were employed along with undersampling of the majority class; for the first iteration each minority class record was added in twice to give a ratio of 2/1 minority to majority; the second where each minority class record was added in three times to give a ratio of 3/1 minority to majority; and the third where each minority class record was added in twice along with a matching random sample of the majority class to give a ratio of 1/1.

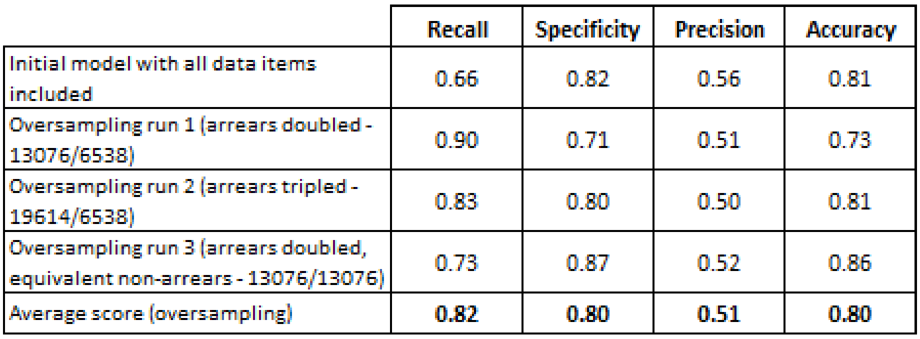
Table 6.5 Oversampling results



	Decision Tree				AdaBoost				SVM				Neural Network			
	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy
Initial model with all data items included	0.74	0.97	0.73	0.95	0.72	0.97	0.75	0.95	0.60	0.97	0.69	0.93	0.57	0.38	0.09	0.40
Oversampling run 1 (arrears doubled - 2/1 ratio)	0.85	0.96	0.69	0.95	0.86	0.95	0.68	0.94	0.87	0.93	0.57	0.92	1.00	0.00	0.10	0.10
Oversampling run 2 (arrears tripled - 3/1 ratio)	0.85	0.96	0.69	0.95	0.87	0.95	0.64	0.94	0.87	0.92	0.55	0.92	0.72	0.39	0.12	0.42
Oversampling run 3 (arrears doubled, 1/1 ratio)	0.85	0.96	0.69	0.95	0.84	0.96	0.69	0.95	0.84	0.94	0.60	0.93	0.40	0.64	0.11	0.62
Average oversampling	**0.85**	**0.96**	**0.69**	**0.95**	**0.86**	**0.95**	**0.67**	**0.94**	**0.86**	**0.93**	**0.57**	**0.92**	**0.70**	**0.34**	**0.11**	**0.38**

Table 6.5 details the scores achieved across each of the four measurements for the models created using oversampling combined with random undersampling. There are considerable improvements in the Recall rate for each of the models, with the exception of one of the Neural Network iterations, though some of the other scores have decreased. Both Specificity and Precision has declined for all of the iterations with random undersampling, with the exception of the Neural Networks where these measures have gone up in the majority of the iterations. The overall Accuracy score for the models in most cases has not changed very much, again with the exception of the Neural Networks, where there is a lot more fluctuation.

Table 6.6 Average results for oversampling



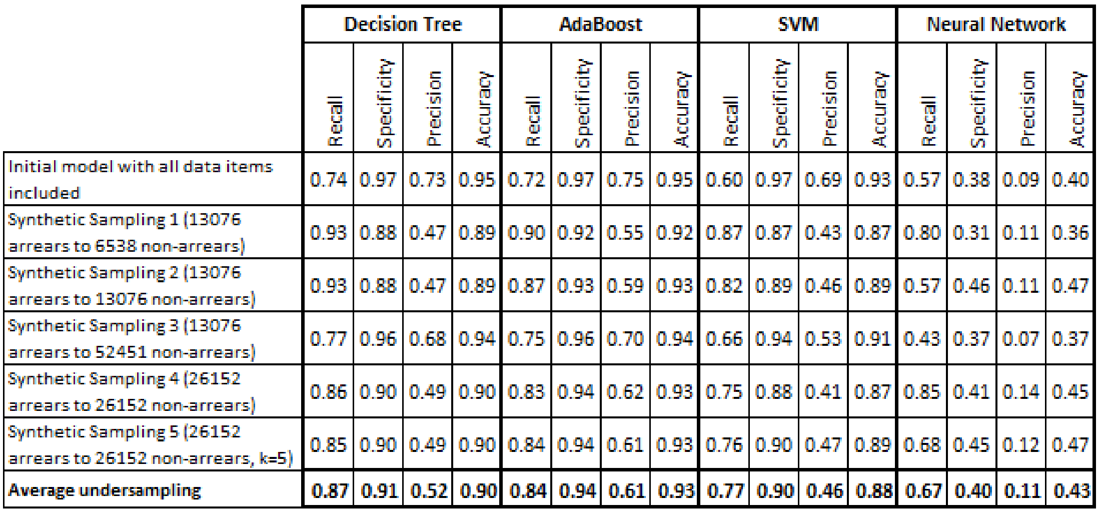
	Recall	Specificity	Precision	Accuracy
Initial model with all data items included	0.66	0.82	0.56	0.81
Oversampling run 1 (arrears doubled - 13076/6538)	0.90	0.71	0.51	0.73
Oversampling run 2 (arrears tripled - 19614/6538)	0.83	0.80	0.50	0.81
Oversampling run 3 (arrears doubled, equivalent non-arrears - 13076/13076)	0.73	0.87	0.52	0.86
Average score (oversampling)	**0.82**	**0.80**	**0.51**	**0.80**

When looking at the average scores from the oversampling, there is a marked increase in Recall score, though again it can be seen that there are slight drops in the Specificity and Accuracy, along with a larger drop in Precision. The different iterations of oversampling performed in different ways for the measures, though the Decision Tree models gave the same overall scores while there was also little variance in the scores achieved with the AdaBoost models.

Synthetic Sampling

The SMOTE algorithm was used to created synthetic minority class data. As discussed earlier, the SMOTE algorithm seeks to create new data items from the minority class that are similar to the existing data items, but are not identical. Five separate iterations of modelling were carried out using synthetic data created. For the first three iterations the number of synthetic data samples produced was 6,538 to match the existing minority class data items and the same synthetic samples were used in each iteration. The difference in the three iterations was the number of majority class records that were used to build the dataset. For the three iterations the number of majority class records that were sampled to complete the dataset were 6,538, 13,076 and 52,541 respectively. This gave ratios of 2/1, 1/1 and 1/4 respectively for the ratio of minority to majority class. The fourth and fifth iteration of synthetic sampling both produced 19,614 synthetic samples. The different between the two iterations was that for the fifth iteration the variable *k* which defines the number of “neighbours” to use in the creation of the synthetic samples, was changed from the default to five to see if this would generate any noticeable difference.

Table 6.7 Synthetic sampling results

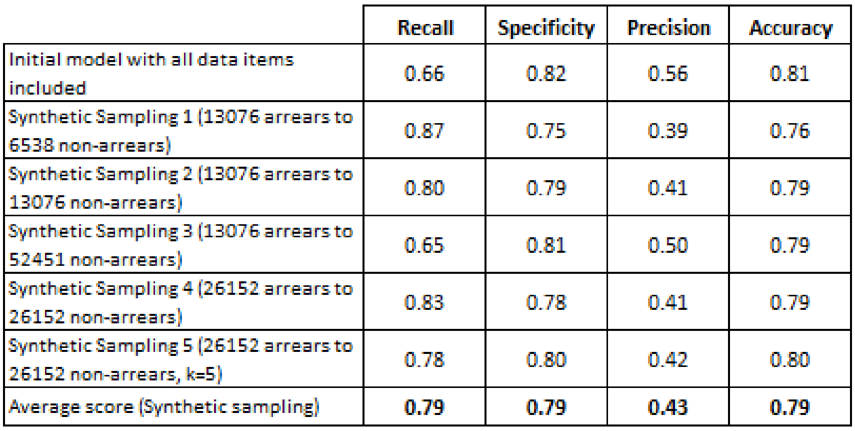


	Decision Tree				AdaBoost				SVM				Neural Network			
	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy
Initial model with all data items included	0.74	0.97	0.73	0.95	0.72	0.97	0.75	0.95	0.60	0.97	0.69	0.93	0.57	0.38	0.09	0.40
Synthetic Sampling 1 (13076 arrears to 6538 non-arrears)	0.93	0.88	0.47	0.89	0.90	0.92	0.55	0.92	0.87	0.87	0.43	0.87	0.80	0.31	0.11	0.36
Synthetic Sampling 2 (13076 arrears to 13076 non-arrears)	0.93	0.88	0.47	0.89	0.87	0.93	0.59	0.93	0.82	0.89	0.46	0.89	0.57	0.46	0.11	0.47
Synthetic Sampling 3 (13076 arrears to 52451 non-arrears)	0.77	0.96	0.68	0.94	0.75	0.96	0.70	0.94	0.66	0.94	0.53	0.91	0.43	0.37	0.07	0.37
Synthetic Sampling 4 (26152 arrears to 26152 non-arrears)	0.86	0.90	0.49	0.90	0.83	0.94	0.62	0.93	0.75	0.88	0.41	0.87	0.85	0.41	0.14	0.45
Synthetic Sampling 5 (26152 arrears to 26152 non-arrears, k=5)	0.85	0.90	0.49	0.90	0.84	0.94	0.61	0.93	0.76	0.90	0.47	0.89	0.68	0.45	0.12	0.47
Average undersampling	**0.87**	**0.91**	**0.52**	**0.90**	**0.84**	**0.94**	**0.61**	**0.93**	**0.77**	**0.90**	**0.46**	**0.88**	**0.67**	**0.40**	**0.11**	**0.43**

For all of the iterations of synthetic sampling with the exception of one iteration for the Neural Networks, the Recall rate was improved upon. This again was at the expense of the Specificity, Precision and Accuracy rates which fell considerably in almost every instance. The average scores achieved per modelling technique from the synthetic sampling show a considerable difference in the classification of the positive class, but with detrimental effects on the classification of the negative class.

From an overall level the average results can be seen to have increased greatly for the Recall rate, but they have changed considerably for the Precision and there is a noticeable difference for both Specificity and Accuracy. The large drop in Precision has caused much of the decrease in overall Accuracy score of the model and this is as a direct result of the synthetic sampling. It would appear that the models are having difficulty differentiating between the synthetic minority class data and the majority class data, as there is a much higher rate of false positives generated. The Specificity rate has also dropped considerably as a result of the false positives increasing. However the Precision rate is skewed much more as a result of the calculation using the true positive rate, which is a much smaller figure than the true negative rate used in the Specificity calculation.

Table 6.8 Average results for synthetic sampling

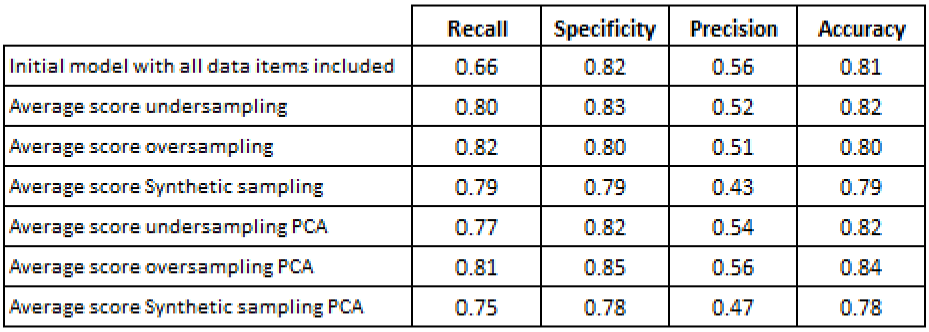


	Recall	Specificity	Precision	Accuracy
Initial model with all data items included	0.66	0.82	0.56	0.81
Synthetic Sampling 1 (13076 arrears to 6538 non-arrears)	0.87	0.75	0.39	0.76
Synthetic Sampling 2 (13076 arrears to 13076 non-arrears)	0.80	0.79	0.41	0.79
Synthetic Sampling 3 (13076 arrears to 52451 non-arrears)	0.65	0.81	0.50	0.79
Synthetic Sampling 4 (26152 arrears to 26152 non-arrears)	0.83	0.78	0.41	0.79
Synthetic Sampling 5 (26152 arrears to 26152 non-arrears, k=5)	0.78	0.80	0.42	0.80
Average score (Synthetic sampling)	**0.79**	**0.79**	**0.43**	**0.79**

Average results achieved across all of the sampling techniques

The average sampling results across all of the models is shown in Table 6.9. The big increases seen in the Recall score for the models with all data items are not replicated in the models using the principal components data.

Table 6.9 Average sampling scores across all techniques



	Recall	Specificity	Precision	Accuracy
Initial model with all data items included	0.66	0.82	0.56	0.81
Average score undersampling	0.80	0.83	0.52	0.82
Average score oversampling	0.82	0.80	0.51	0.80
Average score Synthetic sampling	0.79	0.79	0.43	0.79
Average score undersampling PCA	0.77	0.82	0.54	0.82
Average score oversampling PCA	0.81	0.85	0.56	0.84
Average score Synthetic sampling PCA	0.75	0.78	0.47	0.78

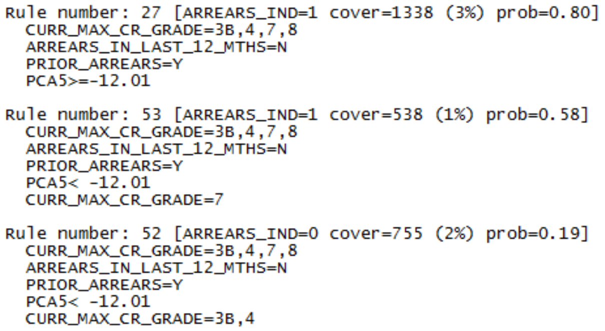
6.2.4 Principal Component Analysis

Principal Component Analysis (PCA) was used to help to reduce the number of variables in the dataset given the large number of numeric variables present (90+). The large number of numeric variables that are in the dataset used for the modelling phase has caused a number of issues. Firstly the sheer number of variables has meant that the modelling phase has taken much longer as each of the algorithms had to process 120+

variables for each of the sixty thousand records in the training/validation dataset. In many cases it took over twenty minutes to build a model for one algorithm, with one of the SVM models taking thirty five minutes to build. The second potential issue with the large number of variables in the dataset was that the models could struggle to identify the key data items from the overall dataset to split the data on.

By using PCA to reduce the number of data items in the dataset the models, the total number of variables in the dataset was reduced by sixty six. This simplified the dataset and allowed the models to be built and trained much quicker. The overall scores of the models were not drastically affected by using the PCA variables instead of the original variables, as 80% of the variance in the data was explained by the principal components generated. Five of the decision tree models built using the principal components instead of the original variables matched the lowest misclassification cost achieved by the other models, so certainly for the decision trees using PCA has had no effect.

The main drawback with using the principal component variables created using the loadings from the PCA is that the level of explainability associated with using the original numeric variables, is lost with the principal components. In Figure 6.2 the principal component PCA5 is used in the three rules, but this variable is composed of all ninety one numeric variables in the original dataset multiplied by the loadings generated by the PCA function. As a result the in-built capability of explaining the rules for a decision tree are lost by using the principal components in the model building.



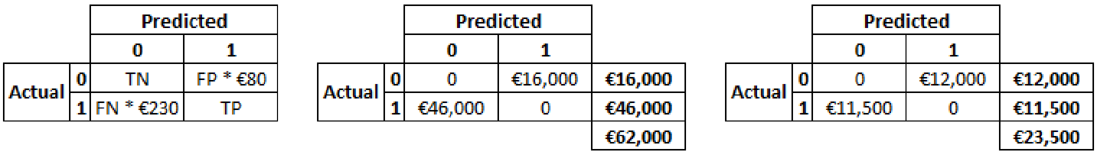
```
Rule number: 27 [ARREARS_IND=1 cover=1338 (3%) prob=0.80]
  CURR_MAX_CR_GRADE=3B,4,7,8
  ARREARS_IN_LAST_12_MTHS=N
  PRIOR_ARREARS=Y
  PCA5>=-12.01

Rule number: 53 [ARREARS_IND=1 cover=538 (1%) prob=0.58]
  CURR_MAX_CR_GRADE=3B,4,7,8
  ARREARS_IN_LAST_12_MTHS=N
  PRIOR_ARREARS=Y
  PCA5< -12.01
  CURR_MAX_CR_GRADE=7

Rule number: 52 [ARREARS_IND=0 cover=755 (2%) prob=0.19]
  CURR_MAX_CR_GRADE=3B,4,7,8
  ARREARS_IN_LAST_12_MTHS=N
  PRIOR_ARREARS=Y
  PCA5< -12.01
  CURR_MAX_CR_GRADE=3B,4
```

Figure 6.1 Principal components used in decision tree rules

6.2.5 Misclassification costs

For this and for most other prediction projects, the misclassification costs associated with the models will be one of the key drivers of which models are chosen as the best models to implement. The misclassification costs will often decide which modelling technique and sampling method has given the best result. For this experiment the misclassification costs will be applied to the confusion matrices created for each run of the models, and where two models have very similar scores the other methods of quantifying the success of the models such as the Recall and Precision rate will be also be utilised. The misclassification costs will be €80 per false positive and €230 per false negative as agreed with the Arrears Support Unit (ASU).

Figure 6.3 shows how the misclassification costs will be applied to the confusion matrix, with an example of the costs for a model where there are 200 false positives and 200 false negatives, and a second example with 150 false positives and 50 false negatives. In most instances it would be expected that the rate of false positives would be higher than the rate of false negatives, as the models should be tweaked so that the rate of false negatives are lower given the associated higher costs.



		Predicted	
		0	1
Actual	0	TN	FP * €80
	1	FN * €230	TP

		Predicted		
		0	1	
Actual	0	0	€16,000	€16,000
	1	€46,000	0	€46,000
				€62,000

		Predicted		
		0	1	
Actual	0	0	€12,000	€12,000
	1	€11,500	0	€11,500
				€23,500

Figure 6.2 Application of misclassification costs to models created

Misclassification Costs for each model produced as part of the experiment

Each of the models produced as part of the experiment will have the misclassification costs applied to the errors produced, so that all models can be compared side by side. For each of the iterations of modelling, the number of errors will be taken from the evaluation of the model against the holdout dataset. The results will be taken for each of the modelling techniques, so that the combination of modelling techniques and sampling techniques can be compared together. It is expected that the sampling techniques will have worked well with some modelling techniques, but not as well with others. All of the comparative results can be seen in Table 6.10.

Decision Trees give the lowest misclassification costs and the most consistent costs of all of the models. The lowest misclassification cost is €63,270 and this cost is achieved by eight of the iterations of the decision trees. This demonstrates that decision trees are able to generalise well, and can produce good scores using a multitude of different sampling techniques as these scores are achieved using both under and over sampling methods, and using the original data items and the principal components identified.

Table 6.10 Model Misclassification costs

	Decision Tree			AdaBoost			SVM			Neural Networks		
Modelling Iteration	**FN**	**FP**	**Cost**	**FN**	**FP**	**Cost**	**FN**	**FP**	**Cost**	**FN**	**FP**	**Cost**
Basic Model with no transactional data	381	123	**€97,470**	333	164	**€89,710**	514	38	**€121,260**	396	197	**€106,840**
Initial model with all data items included	259	267	**€80,930**	273	243	**€82,230**	390	266	**€110,980**	425	5445	**€533,350**
Undersampling run 1 (50/50 majority/minority)	145	377	**€63,510**	146	383	**€64,220**	144	641	**€84,400**	414	5821	**€560,900**
Undersampling run 2 (50/50 majority/minority)	145	375	**€63,350**	151	378	**€64,970**	146	623	**€83,420**	133	3343	**€298,030**
Undersampling run 3 (50/50 majority/minority)	145	375	**€63,350**	148	386	**€64,920**	149	644	**€85,790**	456	5180	**€519,280**
Oversampling run 1 (arrears doubled - 13076/6538)	145	374	**€63,270**	138	404	**€64,060**	136	654	**€83,600**	0	8835	**€706,800**
Oversampling run 2 (arrears tripled - 19614/6538)	145	374	**€63,270**	124	482	**€67,080**	137	690	**€86,710**	277	5389	**€494,830**
Oversampling run 3 (arrears doubled - 13076/13076)	145	374	**€63,270**	156	367	**€65,240**	162	543	**€80,700**	594	3160	**€389,420**
Naïve Bayes all data items	231	2402	**€245,290**									
Synthetic Sampling 1 (13076 arrears to 6538 non-arrears)	72	1020	**€98,160**	100	712	**€79,960**	129	1116	**€118,950**	199	6090	**€532,970**
Synthetic Sampling 2 (13076 arrears to 13076 non-arrears)	131	591	**€77,410**	131	591	**€77,410**	178	950	**€116,940**	421	4745	**€476,430**
Synthetic Sampling 3 (13076 arrears to 52451 non-arrears)	229	349	**€80,590**	247	321	**€82,490**	329	571	**€121,350**	561	5587	**€575,990**
Synthetic Sampling 4 (26152 arrears to 26152 non-arrears)	135	874	**€100,970**	163	505	**€77,890**	244	1064	**€141,240**	143	5210	**€449,690**
Synthetic Sampling 5 (26152 arrears to 26152 non-arrears, k=5)	144	859	**€101,840**	156	526	**€77,960**	232	840	**€120,560**	313	4901	**€464,070**
Initial model with PCA variables added	228	315	**€77,640**	279	268	**€85,610**	490	66	**€117,980**	464	5653	**€558,960**
Undersampling run 1 PCA (50/50 majority/minority	160	372	**€66,560**	146	378	**€63,820**	168	391	**€69,920**	567	4824	**€516,330**
Undersampling run 2 PCA (50/50 majority/minority)	145	374	**€63,270**	133	450	**€66,590**	167	387	**€69,370**	328	5300	**€499,440**
Undersampling run 3 PCA (50/50 majority/minority)	145	374	**€63,270**	153	374	**€65,110**	170	375	**€69,100**	382	5028	**€490,100**
Oversampling run 1 PCA (number of arrears doubled)	145	374	**€63,270**	139	402	**€64,130**	141	475	**€70,430**	392	4581	**€456,640**
Oversampling run 2 PCA (arrears quadrupled)	145	374	**€63,270**	127	465	**€66,410**	131	582	**€76,690**	129	1019	**€111,190**
Oversampling run 3 PCA (2 to 1 arrears to non-arrears) s	145	374	**€63,270**	156	370	**€65,480**	187	339	**€70,130**	374	7043	**€649,460**
Synthetic Sampling 1 PCA (13076 arrears to 6538 non-arrears)	72	1020	**€98,160**	118	676	**€81,220**	158	465	**€73,540**	414	5530	**€537,620**
Synthetic Sampling 2 PCA (13076 arrears to 13076 non-arrears)	72	1020	**€98,160**	130	602	**€78,060**	204	402	**€79,080**	416	6236	**€594,560**
Synthetic Sampling 3 PCA (13076 arrears to 54541 non-arrears)	203	622	**€96,450**	245	365	**€85,550**	380	216	**€104,680**	510	5867	**€586,660**

The AdaBoost algorithm was close to achieving similar costs to the decision tree models in some of the iterations, and the lowest cost was €63,820 which is only just above the best cost for the decision trees. As with the decision tree models, in some cases the AdaBoost model does well on the false negative costs, but it is let down by the costs associated with the false positives – especially with the synthetic sampling iterations.

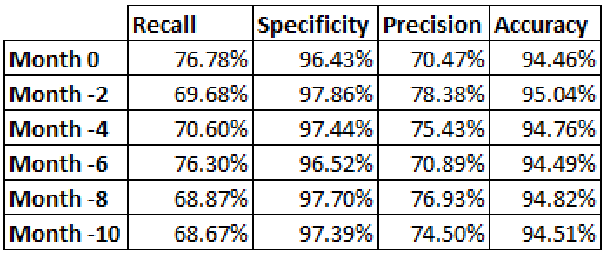
With the exception of a small number of models that performed relatively closely to the decision trees and AdaBoost models, most of the SVM models had a much higher misclassification cost associated. In some cases this was due to the false positives, but for a number of the iterations the false negatives were quite high. The Neural Network models performed very badly in almost every instance from a misclassification cost perspective. For one of the iterations the model correctly predicted all positive records correctly but incorrectly classified all of the negative class bar four records. The resultant costs for this model were the highest of all the models at €706,800 which is a very long way from the worst scores for the other algorithms.

6.2.6 Predicting arrears in advance

The goal of predicting the mortgages likely to go in to arrears in the following months is to give Lender A at least a month or two to be able to intervene with the borrowers identified as at risk. This additional time should theoretically allow the lender to put a process in place that will ultimately lead to the arrears being avoided, or at the very least lessened. For the purposes of predicting the arrears in advance the model was run at two month intervals from the date immediately prior to arrears to ten months prior to the arrears.

Each of the models created at the bi-monthly intervals have been scored using the measurements already discussed, and the results are show in Table 6.11. Month 0 is taken as the baseline month to compare each of the other prediction models against and each of the models are built against the full dataset with all numeric variables present.

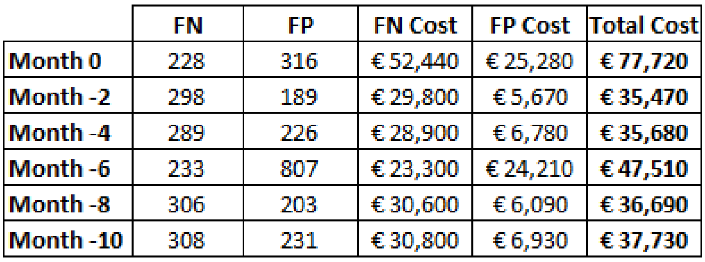
Table 6.11 Advance model prediction scores



	Recall	Specificity	Precision	Accuracy
Month 0	76.78%	96.43%	70.47%	94.46%
Month -2	69.68%	97.86%	78.38%	95.04%
Month -4	70.60%	97.44%	75.43%	94.76%
Month -6	76.30%	96.52%	70.89%	94.49%
Month -8	68.87%	97.70%	76.93%	94.82%
Month -10	68.67%	97.39%	74.50%	94.51%

For all of the months with the exception of Month -6 there is a noticeable drop in the Recall rate, and the drop from month 0 to Month -2 is quite steep. The subsequent decreases noted in the other months could almost be compared to a linear decrease (with the exception of Month -6). There is very little difference in the Specificity rate when looking at the different months, though there is a noticeable drop in Month -6. When looking at the number of false negatives and positives in Table 6.12 it is easier to identify the differences between the models, as the numbers change substantially across the different models. At a first glance when looking at the Recall rate of Month -6 it would appear that the model operates at a similar overall accuracy as the Month 0 model, but this is not the case. The false negatives are certainly comparable, but the number of false positives has increased drastically for Month -6. Even taking in to account the sliding scale of the misclassification costs for Month -6, the model misclassification costs of €47,150 are quite high compared to the other iterations. Looking at the results for the other models it would appear that either the model used for Month -2 or Month -4 could be implemented without much loss of accuracy. Both of these models have an increased level of false negatives, though the number of false positives has decreased considerably. Weighing up the two options, the ideal model would be the Month -4 model as this would give the lender sufficient time to engage with the borrower.

Table 6.12 Advance model misclassification costs



	FN	FP	FN Cost	FP Cost	Total Cost
Month 0	228	316	€ 52,440	€ 25,280	€ 77,720
Month -2	298	189	€ 29,800	€ 5,670	€ 35,470
Month -4	289	226	€ 28,900	€ 6,780	€ 35,680
Month -6	233	807	€ 23,300	€ 24,210	€ 47,510
Month -8	306	203	€ 30,600	€ 6,090	€ 36,690
Month -10	308	231	€ 30,800	€ 6,930	€ 37,730

Sampling techniques were not applied to the data used for the advance model predictions. However, it is assumed that comparable gains would also be experienced if sampling was applied to this data, as was seen with the general prediction model.

6.3 Interpretation of Results

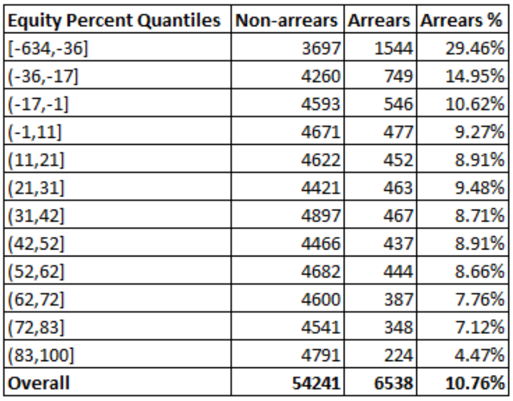
The research on mortgage arrears and prediction models identified some of the characteristics of borrowers who are difficulty such as households with dependent children being much more likely to go in to arrears (McCarthy, 2014). A field was included in the dataset that indicated if the borrowers had dependent children, but it appears that this is not a useful variable for predicting arrears as it was not used in building any of the models. It could be that the data used by McCarthy pointed towards dependent children being associated with mortgages in arrears simply because the vast majority of borrowers with arrears had children, rather than dependent children being a causal factor in the arrears.

Unemployment or lack of regular employment was also noted as factor in mortgage arrears though this could not be verified by this experiment. There is data available on employment status for all personal customers in Lender A as well as salary amount, but in many cases this has not been recently updated. Many borrowers who are in difficulty may not have engaged with the bank in a number of years given the nature of how banking has changed in the last few years with fewer and fewer customers actually entering branches to carry out their transactions. As a result the employment and salary figures held for many mortgage holders is out of date and should be updated. In the case of borrowers who are already engaged with the lender the information is likely to be up to date, but for borrowers not engaged it is likely this information is not fully up to date. Without having valid and up to date employment information on all borrowers it is not possible to see what role unemployment may have in the mortgage arrears being experienced in Lender A.

The “Double Trigger” cause of mortgage arrears and default was discussed by many researchers, where they felt that an underlying issue such as negative equity by itself did not cause arrear. They did argue though that in the case of mortgages in negative equity if a borrower suffered a second major “trigger” such as a loss of income or illness, then the borrower is much more likely to miss payments on their mortgage and go in to arrears. Without having information relating to events like job losses or illness (except in the situation where a borrower is engaged in a formal process with the lender) it is very hard to identify the double trigger effect. This invariably does not mean that the double trigger effect does not exist in the mortgages in the sample in difficulty, but a greater depth of analysis would be necessary to review the mortgages in arrears to identify if there is a double trigger.

From the standpoint of negative equity it would appear that the ratio of mortgages in arrears that have negative equity is much higher than for the mortgages with positive equity. There is a higher level of mortgages in arrears that have positive equity the mortgages with negative equity, but when looking at the ratios it shows a different picture. 22.62% of the mortgages in the training sample that are in negative equity are in arrears, while only 8.87% of the mortgages with positive equity are in arrears. This is a very high level assumption to make as there will be both mortgages in negative equity with no arrears as well as mortgages with positive equity that are in arrears. Figure 6.6 shows the breakdown of the total percentage of mortgages in arrears when the equity percent is divided into twelve quantiles.

Table 6.13 Mortgage arrears by equity percent quantiles



Equity Percent Quantiles	Non-arrears	Arrears	Arrears %
[-634,-36]	3697	1544	29.46%
(-36,-17]	4260	749	14.95%
(-17,-1]	4593	546	10.62%
(-1,11]	4671	477	9.27%
(11,21]	4622	452	8.91%
(21,31]	4421	463	9.48%
(31,42]	4897	467	8.71%
(42,52]	4466	437	8.91%
(52,62]	4682	444	8.66%
(62,72]	4600	387	7.76%
(72,83]	4541	348	7.12%
(83,100]	4791	224	4.47%
Overall	**54241**	**6538**	**10.76%**

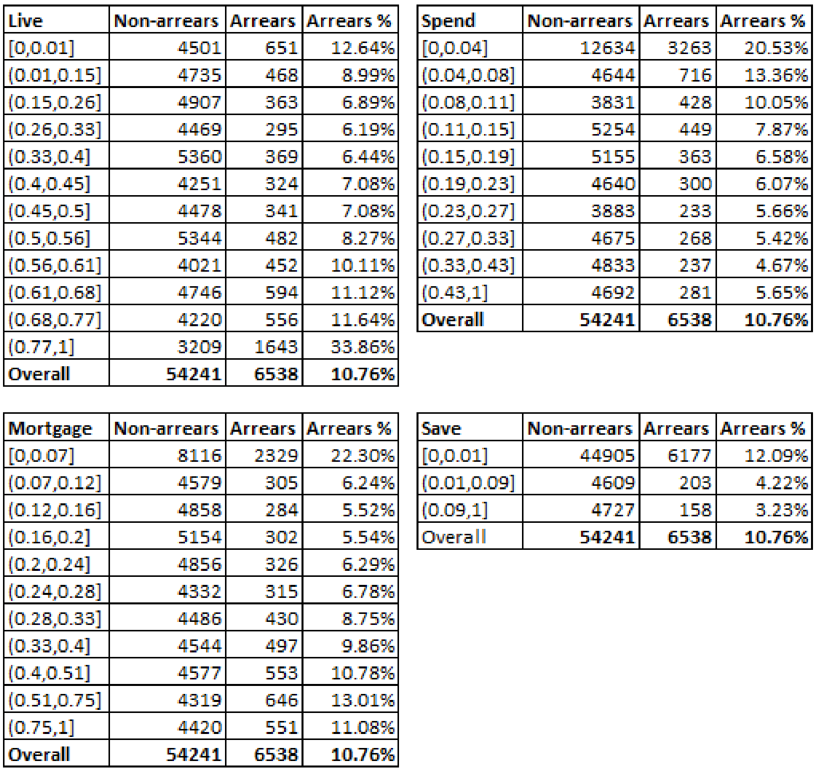
While there is a linear pattern to the rate of arrears which drops sharply as the equity held on a property gets closer to 100%, only two of the quantiles have a rate of arrears that is above the overall rate of arrears. In the lowest band of negative equity the rate of arrears is almost 30%, which would lend credence to the literature of (Bhutta et al., 2010) who stated that most borrowers will default on their loan when the negative equity rate gets to a certain point. Again though this does not help to validate the double trigger theory without any evidence showing the second trigger that actually causes the arrears.

Arrears are a recurring event for many borrowers, who go in to arrears on their mortgage and are able to catch up on their payments, but once again fall back in to the cycle of arrears at a later date. This is seen in the models by the importance of the fields that indicate where a mortgage has previously been in arrears. This is in part due to the policies and actions of Lender A. In the past Lender A did not have sufficient manpower to be able to address all of the arrears cases sufficiently such was the rate of expansion of mortgages in difficulty. The staff dealing with mortgage arrears were stretched and unable to cope with the volume of arrears they were facing.

The headcount has been increased considerably in the Arrears Support Unit and they now have sufficient staff to deal better with the arrears situation. This should allow more mortgages where the borrower is in difficulty, to be properly resolved so that the borrower should not once again dip in to the arrears cycle unless an unforeseen circumstance occurs. In some instances where the borrower will be unable to afford to meet the required payment obligations, under the Mortgage Arrears Resolution Process (MARP) the bank will have to offer the borrower a long term solution. In some cases the solution may be a debt write-down, a split mortgage or potentially even a sale for loss. However, in most cases it is likely that it will be reduced payments or a payment holiday for a specified period of time.

The transaction level data that was supplied to the model in the form of the Spend/Save/Live/Mortgage categories certainly shows differences in borrowers who are in arrears and those who are not in arrears as shown in Figures 5.7 and 5.8. However, just because the level of spending of a certain borrower is in a pre-defined range, this does not automatically mean that they will either default or not. Figure 6.7 shows the high level transaction categories divided in to quantiles with the percentage of each quantile that are in arrears.

One of the most noticeable characteristics is that in the Save category, 83% of the mortgages are in the first quantile which ranges from 0 to 0.01 total Save spending. Of these mortgages 12.09% are in arrears, but this means that 88% of the mortgages with little or no saving on a regular basis are not in arrears. From Figures 5.7 and 5.8 it would appear that borrowers who are not in arrears save a considerable amount while borrowers it arrears do not save at all. Using the figures from the Save quantile there are nearly 39,000 borrowers who are not in arrears who are not regularly saving either. This would point to the fact that borrowers who don't save regularly do not necessarily go in to arrears. However, any borrower who goes in to arrears and has not been saving regularly up to that point will likely find it very difficult to extricate themselves from arrears. It would therefore seem imperative that borrowers would be encouraged to save on a regularly basis for a "rainy day".



Live	Non-arrears	Arrears	Arrears %
[0,0.01]	4501	651	12.64%
(0.01,0.15]	4735	468	8.99%
(0.15,0.26]	4907	363	6.89%
(0.26,0.33]	4469	295	6.19%
(0.33,0.4]	5360	369	6.44%
(0.4,0.45]	4251	324	7.08%
(0.45,0.5]	4478	341	7.08%
(0.5,0.56]	5344	482	8.27%
(0.56,0.61]	4021	452	10.11%
(0.61,0.68]	4746	594	11.12%
(0.68,0.77]	4220	556	11.64%
(0.77,1]	3209	1643	33.86%
Overall	**54241**	**6538**	**10.76%**

Spend	Non-arrears	Arrears	Arrears %
[0,0.04]	12634	3263	20.53%
(0.04,0.08]	4644	716	13.36%
(0.08,0.11]	3831	428	10.05%
(0.11,0.15]	5254	449	7.87%
(0.15,0.19]	5155	363	6.58%
(0.19,0.23]	4640	300	6.07%
(0.23,0.27]	3883	233	5.66%
(0.27,0.33]	4675	268	5.42%
(0.33,0.43]	4833	237	4.67%
(0.43,1]	4692	281	5.65%
Overall	**54241**	**6538**	**10.76%**

Mortgage	Non-arrears	Arrears	Arrears %
[0,0.07]	8116	2329	22.30%
(0.07,0.12]	4579	305	6.24%
(0.12,0.16]	4858	284	5.52%
(0.16,0.2]	5154	302	5.54%
(0.2,0.24]	4856	326	6.29%
(0.24,0.28]	4332	315	6.78%
(0.28,0.33]	4486	430	8.75%
(0.33,0.4]	4544	497	9.86%
(0.4,0.51]	4577	553	10.78%
(0.51,0.75]	4319	646	13.01%
(0.75,1]	4420	551	11.08%
Overall	**54241**	**6538**	**10.76%**

Save	Non-arrears	Arrears	Arrears %
[0,0.01]	44905	6177	12.09%
(0.01,0.09]	4609	203	4.22%
(0.09,1]	4727	158	3.23%
Overall	**54241**	**6538**	**10.76%**

Figure 6.3 Spend/Save/Live/Mortgage category quantile arrears percentages

In both the Live and Mortgage categories the rate of arrears is higher at both the high and low end of the quantiles with the rate of arrears falling off in the middle. For both of these categories it would make sense that the rate of arrears is higher when the rate of spending is higher as this would signify a higher level of necessary spend on food, bills, utilities and mortgage etc. The lower end of the spending though does not make sense though especially for the mortgage quantile where 22.3% of mortgages with a spending of between 0 and 0.07 are in arrears. In the Spend category the two lowest quantiles have the highest rate of arrears which is consistent the theory that borrowers in difficulty will likely be spending much less on discretionary spend than borrowers who are not in difficulty.

The experiment carried as part of this research shows that it is possible to build a prediction model that would be able to accurately predict mortgage arrears. Further work is required though to be able to work out a solution to the imbalance in the classes, to ensure that the minority class is predicted as accurately as possible without sacrificing the accuracy of the predictions for the majority class. Given the imbalance in the dataset this is likely the most important facet of the model to be worked on.

6.4 Evaluation of Tools Used – R and Rattle

R is a very powerful tool that allows both graphical and statistical analysis to be carried out on the data as well as the ability to perform data mining functions. As a result of the capabilities of R the vast majority of this experiment has been carried out within the R environment, with the exception of extracting the data from the warehouse tables, and manipulating the data to get it in to a usable format for the modelling phase. Using R and Rattle as the main modelling tools for this experiment has had numerous advantages, but there have also been a number of disadvantages and drawbacks.

Some of the many advantages of R such as the speed of loading new data were utilised during the experiment, as there were a large number of iterations of the data which had to be loaded each time to R when the data was changed. It is possible to load the data directly in to Rattle using the GUI which adds the data to R anyway, but it proved easier to load the data to R first and then pick the R dataset as the basis for the modelling within Rattle and define the variables there if necessary.

Due to the expansive list of functions and packages available in R and Rattle, it was not necessary to use a separate tool to explore and visualise the data, so this certainly cut down on some effort required to load the data to a different application and carry out the exploration there. Within the Rattle application itself though there were some limitations to what could be done with visualizations, so much of the exploration and graphs was done purely using the R console.

The R statistical language came in to existence long before there was any mention of "Big Data", and was initially used mainly as a statistical tool for measuring and manipulating the output of experimental data. R was designed to process data in memory, so with large datasets this can cause some issues. While the dataset used as part of this experiment is not quite on the scale of Big Data, nevertheless in the training/validation dataset there are 60,779 records with one hundred and twenty columns of data (excluding the principal components generated). Processing this amount of data in R can cause problems as the memory utilised by the application can quickly be consumed, causing the application to perform poorly and ultimately crash.

Memory issues were particularly evident when utilising the SVM modelling technique, as the machine being used had its RAM capacity capped due to network restrictions within Lender A. In most cases when using the SVM technique the computer had to be re-started and only Rattle could be opened and the SVM processed prior to opening any other applications, or carrying out any other modelling within Rattle. This meant that multiple runs of the modelling were necessary in order to be able to compare different modelling techniques – especially when using the full dataset. This was later addressed by using a different computer not on the internal network with 6GB of RAM and a 64-bit processor, but all the modelling had to be repeated to ensure the results were correct.

The Rattle application seemed to crash on quite a regular basis without warning, and would abort the R session, as well as losing any data that was contained within the application, but not saved. This was remedied somewhat by loading the data separately in R and then processing in Rattle, but it certainly slowed down the whole process.

Almost all of the required functions for the processing and modelling of the data for this experiment such as PCA were available within Rattle. This allowed PCA to be carried out on the data so that the number of numeric variables in the dataset could be reduced. However, given that the PCA tool is available in Rattle it would be very useful if the output of the PCA could be utilised and applied to the dataset to create the principal components within Rattle/R. This would seem to be a shortcoming within the application, as the loadings for each component/variable are available within the application, but there is no apparent way to apply them to the original data to get the principal components. As a result all of the PCA loadings were applied to the original data items within the data warehouse, to create each of the 35 principal components which required a large amount of SQL code, and left a lot of room for errors in the calculations. If the results of the PCA could be directly applied to the data within Rattle/R then this would certainly be a useful piece of functionality, and would cut down on the amount of manual work required as well as the potential for error.

Overall, despite some of the issues encountered the combination of R and Rattle used for this experiment proved to be a successful one. It is possible though that some of the more commercially established tools may have been more successful at processing the large amount of data without the risk of crashing so often.

6.5 Overall Evaluation of Experiment

The experiment to design and build the prediction model that would predict mortgage arrears was completed successfully and produced a model which is able to predict arrears with a high level of accuracy. The standard predictive model that tries to forecast which mortgages are going in to arrears was very effective; with a relatively high Recall rate of over 85% and good scores across the other measurements. The model built to predict arrears a number of months in advance also performed well, and gave good predictions at both two and four months prior to arrears. In practice the model to predict arrears a number of months in advance would need some further work to ensure that concept drift does not mean that the model will become redundant after a number of months.

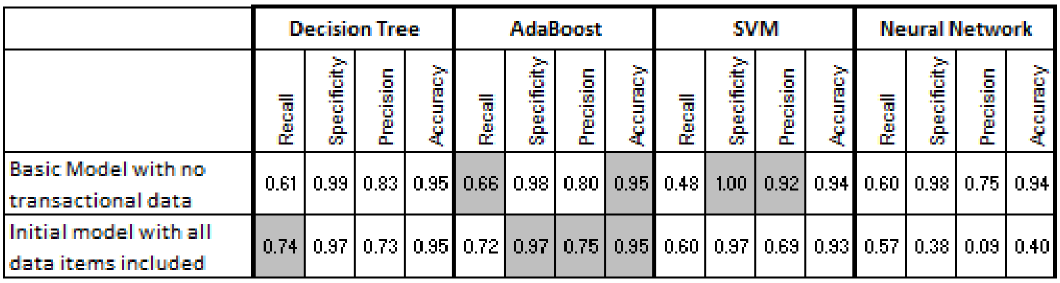
Concept drift in fact would be important for any prediction model put in place by Lender A. A mortgage arrears prediction model would likely need to be re-trained at regular intervals to ensure it is still functioning correctly. For the case of the models that predict arrears a number of months in advance concept drift is more of an issue. The reason for this is that the output of the models will be used to contact borrowers to try to ensure that they do not go in to arrears in their mortgage. As a result of the lender contacting the customers, the behaviour of borrowers would likely change and it would prove more difficult to actually predict arrears.

Concept drift would likely also be an issue given wider macroeconomic factors. The rate of mortgage arrears in Ireland was relatively low before the onset of the financial crisis, which caused the rate of arrears to increase rapidly. In the same way the return to a more “normal” functioning economy should see a drop in the number of mortgages in arrears as unemployment numbers decrease and wages once again increase. As a result the overall economy is likely to have an effect on any arrears prediction models, and this would need to be taken in to account when re-training the models.

As the dataset used for building the model has a much higher ratio of good mortgages a number of approaches were taken to try to address the class imbalance. Some of the approaches such as random undersampling and oversampling were more successful than others, as synthetic sampling did not show very good all over results. Synthetic sampling showed good promise with the Recall rate, and it had the lowest number of false negatives of any of the other approaches. However, this increase in the detection of the minority class was at the expense of the majority class. Though not explored as part of the experiment, it is possible that an ensemble model would be better suited to addressing the class imbalance. An example application could be a clustering algorithm that would seek to cluster together the borrowers with the output feeding in to the classification algorithm to predict the arrears.

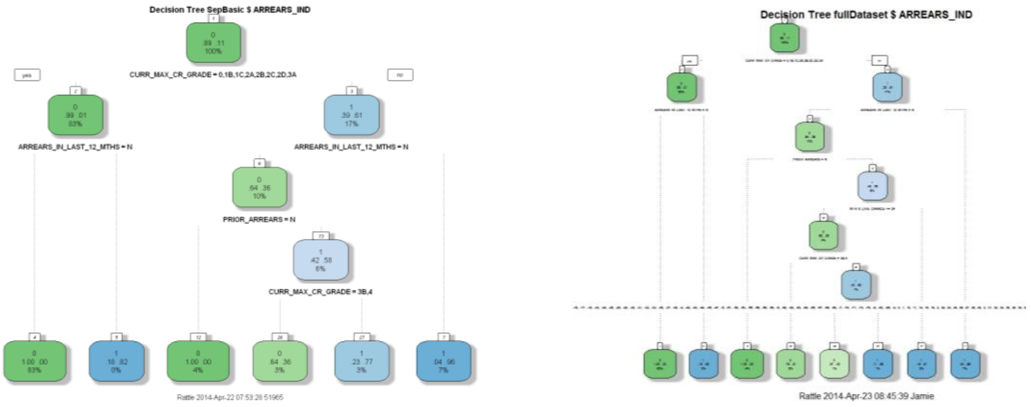
While the models created are capable of predicting arrears with a high level of accuracy, another objective of the experiment was to ascertain if the transaction level and negative equity data would improve on the predictive capability of the model. The initial model created without this data scored relatively well on all measures for the AdaBoost model as can be seen in Table 6.14. However, when this additional data was added to the model the vast majority of the scores increased with the exception of the Neural Network scores. Though there were decreases in the Specificity and Precision for most of the models, the Recall rate showed a very good improvement with a 13% increase for the Decision Tree model alone.

Table 6.14 Initial model vs. model with transaction data



	Decision Tree				AdaBoost				SVM				Neural Network			
	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy	Recall	Specificity	Precision	Accuracy
Basic Model with no transactional data	0.61	0.99	0.83	0.95	0.66	0.98	0.80	0.95	0.48	1.00	0.92	0.94	0.60	0.98	0.75	0.94
Initial model with all data items included	0.74	0.97	0.73	0.95	0.72	0.97	0.75	0.95	0.60	0.97	0.69	0.93	0.57	0.38	0.09	0.40

The initial Decision Tree model created using all of the available data made use of just one of the transaction level data items, but also took in to account the equity percent data for the splits. Both the basic Decision Tree splits and the Decision Tree splits for the full dataset can be seen in Figures 6.4(a) and Figure 6.4(b) respectively. The Decision Tree for the dataset including the transaction level and equity percentage data is slightly more complex, but it is not an overly complex tree as far Decision Trees go.


Decision Tree SepBasic $ ARREARS_IND
1
0
.89 .11
100%
yes
CURR_MAX_CR_GRADE = 0,1B,1C,2A,2B,2C,2D,3A
no
2
0
.99 .01
83%
ARREARS_IN_LAST_12_MTHS = N
3
1
.39 .61
17%
ARREARS_IN_LAST_12_MTHS = N
6
0
.64 .36
10%
PRIOR_ARREARS = N
13
1
.42 .58
6%
CURR_MAX_CR_GRADE = 3B,4
4
0
1.00 .00
83%
5
1
.18 .82
0%
12
0
1.00 .00
4%
26
0
.64 .36
3%
27
1
.23 .77
3%
7
1
.04 .96
7%
Rattle 2014-Apr-22 07:53:28 51965
Decision Tree fullDataset $ ARREARS_IND
Rattle 2014-Apr-23 08:45:39 Jamie

(a) **(b)**

Figure 6.4 Decision Tree rules

The AdaBoost model improved its Recall score, though only by 6% which is relatively small compared to the increase given by the Decision Tree. The variable importance diagram for the model shows there are eleven transaction level fields from the total of thirty, which are derived from the transactional data. The AdaBoost model in this instance did not make use of the equity percentage data. The Decision Tree was able to generalise with less data items while using the equity percentage, but the AdaBoost model has iterated through the data using more of the transactional data to try and correct the misclassified items. Ultimately though in this instance the AdaBoost model was not able to correctly classify as many minority class records as the Decision Tree.

The scores of the Neural Network decreased dramatically. Neural Networks apply weights to each of the fields to achieve an end result, and the additional ninety numeric fields would have made the model very complex with a huge number of weights. The initial Neural Network scores were relatively good but this was with approximately thirty fields, so the model was much less complex.

6.6 Conclusions

This chapter has presented an evaluation of the results achieved in the experimental phase of this project. Each of the models have been evaluated based on a number of factors including the four main measurements used for success of the models – Recall, Specificity, Precision and overall Accuracy as well as by using the misclassification costs.

The chapter documented the different sampling techniques applied to the data to address the imbalance in the dataset as well as the application of Principal Component Analysis to the dataset to decrease the number of variables required. An appraisal of the both of the main tools used – R and Rattle was presented, listing both the advantages and disadvantages of both. The chapter concluded with an overall interpretation of the results and evaluation of the experiment completed.

The results have shown that the model to predict arrears is capable of predicting the arrears with a high level of accuracy for the target class. The model built to predict arrears a number of months in advance also performed well for both two and four months in advance of the arrears occurring on the mortgages.

The misclassification costs applied to the models have shown quite a difference in costs associated across the different modelling techniques. While the majority of the costs associated with the misclassification costs are essentially operational costs, the overall misclassification costs still allow a useful comparison of each of the models. The models that have scored best on both Recall and Precision have also produced the lowest misclassification costs, with a number of models giving the same costs.

Due to the imbalances nature of the dataset, a number of different sampling techniques were applied to the data to improve the predictions on the target class. Sampling was shown to elicit a good response in improving the recall score of the models, but in many cases it has meant that the Specificity and Precision scores have decreased. Further work is required on implementing different sampling techniques or potentially modifying some of the model parameters to try to produce a model that will score highly on both the minority and majority classes.

7. CONCLUSIONS AND FUTURE WORK

7.1 Introduction

This chapter concludes the dissertation, and will summarise the key points of the research. The introduction will review the reason for the research along with the challenges encountered and the benefits of the experiment. An overview of the research definition and objectives will be covered along with the contributions to the body of knowledge.

An evaluation of the experiment along with an evaluation of the overall research will be presented along with any limitations to the research or experiment. Ideas and areas of interest for future work and research will be highlighted followed by the concluding remarks.

7.2 Research Definition and Research Overview

The research carried out as part of this dissertation was to review the current literature on mortgage arrears and default, as well as the available literature on prediction models. This research was used to help design and implement the experiment to build the predictive models, as well as to evaluate the performance of the models. The models created were built so that they would be able to predict arrears in the mortgage dataset with a high level of accuracy, while reducing the misclassification costs as much as possible. As part of the research the following objectives were achieved:

- Reviewed current literature on mortgage arrears/default and data mining techniques to find the current algorithms used in arrears/default prediction models
- Analysed all of the available data to identify suitable data and derivations for the modelling
- Designed and built a prediction model capable of predicting what mortgages are likely to go in to arrears
- Used a dataset of previously unseen data to test the accuracy of the model
- Employed a number of sampling techniques to endeavour to achieve a balance in the training/validation dataset

- Evaluated each of the models to see which give the best predictions as well as the lowest overall misclassification costs
- Evaluated the overall success/failure of the experiment and dissertation

7.3 Contributions to the Body of Knowledge

The following findings and results can be considered to be contributions to the body of knowledge achieved as part of this dissertation:

- Demonstrated that using derived transactional level data in the prediction model helps to improve the predictive capability of the model.

- Shown that adding equity percentage data to the prediction model helps to improve the ability of the model to split the data, but further research is required to ascertain the action of negative equity in the "Double Trigger" theory.

- This work has established than when working with an imbalanced dataset, applying sampling techniques to the training/validation dataset can help to increase the overall accuracy of the predictions for the minority class.

- Despite the lack of academic literature showcasing the power and accuracy of Decision Trees in predictive models, this experiment has shown that Decision Trees are certainly a viable alternative to some of the more complex models.

7.4 Experimentation, Evaluation and Limitations

This experiment set out to build a model that was capable of predicting arrears in the mortgage dataset sourced from Lender A. The objective of the experiment was first to build a working model that was capable of predicting arrears; and then build on this baseline to incorporate transactional data in the hope that this would improve the accuracy and predictive capability. Investigating the usefulness of negative equity as an input to the model was an additional objective of the research. A further model was developed that would be able to predict arrears a number of months in advance of the arrears actually happening on the mortgages. Much of the work of building the models was trying to address the imbalance in the dataset, as it is heavily weighted towards mortgages that are not in arrears.

The initial model built without the transaction data was relatively successful at predicting arrears with a good accuracy at predicting the mortgages that were going to fall behind on their payment obligations. When the transactional data was added in to the model building process the rate of success at predicting the arrears went up again which would indicate that the transactional level data does indeed help to improve the predictive capability of the model. The sampling techniques applied to the data to remedy the imbalance in the dataset had varying levels of success with some such as the synthetic sampling giving a very good score on the positive target class, but a much diminished score on the negative class. The simple random undersampling and random oversampling did achieve good scores without being too difficult to implement.

Overall the experimentation carried out as part of this dissertation has been successful at achieving its goal of building a model that will predict arrears. The experiment was also successful at testing whether or not both transaction level and negative equity data can help to improve the model, as both were successfully used in the models produced. The model built to predict arrears a number of months in advance produced acceptable results, though a refined implementation of this model would require further work.

The models produced as part of this research helped solving an acute and real business problem within Lender A. While there are models in place to predict what mortgages are going to go over arrears of ninety days, little is known about what mortgages are going to miss a payment in the next month and decline into arrears. Both the Head of Data and Head of Data Innovation within the Data Domain have been very positive about the models produced and the results achieved. The representatives from both the Financial Services Group (FSG) and Arrears Support Unit (ASU) were both also very pleased with the outcome and have suggested further work that could extend the project.

Much of the data that was used to build and train the model was based on transactional data that was taken from all of the accounts of the borrowers associated with the mortgage. In some cases a borrower may have their mortgages with Lender A, but their current account and credit cards may be with another institution. If a borrower has their current account and credit card with another bank then it will not be possible to see all of their transactions, and hence the Spend/Save/Live/Mortgage transaction figures will not be totally correct. In the majority of cases a borrower will have both their current account and mortgage with the same bank, so this will not be an issue for most of the mortgages in the sample.

While the models created as part of this experiment seek to predict what mortgages are going to go in to arrears, the focus of the research and experiment were never to work out why the mortgages are going in to arrears. Some of the research covered documented reasons for borrowers getting in to arrears, but this would need to be the subject of further research and analysis to discover the underlying reasons for the mortgages getting in to difficulty. Some of the research covered made reference to strategic defaults by borrowers who are able to pay their mortgage but choose not to. This was beyond the realms of this experiment, though it could be covered in a future extension of this work.

7.5 Future Work and Research

After engaging with FSG and ASU within Lender A who deal with borrowers in difficulty and mortgage arrears, they have taken a greater interest in what can be done with the data available in the data warehouse. There are a number of areas they would like to investigate further to see how they can get a better understanding of what is happening with the mortgages in arrears and how they can prevent and manage the arrears.

One of the key areas to investigate is how to predict what mortgages are going to go in to arrears that have not previously been in any difficulty. This is quite similar to the model that has already been built, though some adjustments will have to be made. Two of the important fields used in splitting the data in the decision trees were whether or not the mortgage had previous arrears and whether or not it had arrears in the previous twelve months. Both of these fields will not be available in the new model, so additional data may be necessary to allow for a good model to be created.

Part of the work that is done in both FSG and ASU as well in other areas of Lender A is to monitor the wider economic climate for changes that will affect how the bank does business. In the current financial climate they are focused on monitoring the economy for signs that consumer spending and jobs are returning to close to pre-financial crisis levels. If they are able to tell on a wide scale if the economy is improving then this can have a big impact on the policies of the lender, both in terms of how they are treating new business as well as how they are dealing with existing loans and arrears. As part of this work FSG would like to extend the analysis carried out on the transactions to give a macroeconomic view of the economy based on the transaction habits of the customers in Lender A. This macroeconomic view will use a combination of the rules from the Money Manager application and the extraction and classification rules from this experiment to classify all transactions for personal customers. FSG would like to use this for both analysis of customers currently managed by them, as well as a time-series analysis of the wider economy over the past number of years.

Following on from the prediction of arrears and wider macroeconomic analysis, Lender A would also like to create intra-month forecasting that will monitor and predict the current state of arrears in a month as well as try to predict what the likely end of month situation will be. In a large percentage of cases borrowers go in to arrears when their payment is due, but make up this payment or “self-cure” within the month. There are many reasons for this including irregular salary payments in to the funding account for the mortgage, or simply that some borrowers do not have a regular payment such as a direct debit setup on the mortgage. FSG would like to be able to analyse the data on mortgages in great detail so that they know the exact position of the arrears on a daily and weekly basis and know how many mortgages are likely to go in to arrears, as well as how many are likely to self-cure. Part of the analysis will be to look at the mortgages in two of the arrears categories – 60-90 days past due (2-3 mortgage payments missed) and those in the 90+ days past due (at least 3 mortgage payments missed). Given the additional capital requirements needed for mortgages in the 90+ days past due category they would like to be able to predict what mortgages are going to move in to this category during the month as well as what mortgages are likely to move from the 90+ category to the 60-90 days category. This analysis will require elements of the prediction model built for both new and existing arrears, as well as a good understanding of the payment details for the mortgages.

The approaches taken to address the imbalance in the dataset were successful to a certain degree, but did not ultimately provide a solution where both the false negatives and positives were kept to a minimum. There was some degree of success in the experiment with the SMOTE synthetic sampling as some of the false negative rates were decidedly lower, but at the expense of the false positive rate. The Borderline-SMOTE was a technique that was discussed in the literature review, but no implementation of this for R could be found to be applied to the dataset. Further research on both this and another of the synthetic sampling approaches – ADASYSN, could be carried out to ascertain if either of these approaches would be able to provide a better resolution to the imbalance.

If the prediction model is implemented in Lender A and borrowers identified as likely to go in to arrears contacted and ultimately the arrears stopped, the model will need to be updated on a regular basis. This is necessary as the behaviour and activity of borrowers will have changed – especially if the model to predict arrears a number of months in advance is used. This is in line with the concept drift theory that would suggest that the target is likely to keep changing as the model progresses and matures.

Further future work could also be carried out to ascertain what portion of the arrears are by strategic defaulters. Many researchers state that a relatively large portion of arrears and defaults are from borrowers strategically defaulting. By analysing the borrowers in arrears in great detail it should be possible to work out if there are any strategic defaulters in the mortgage book of Lender A, so that these borrowers could be engaged with.

7.6 Conclusions

This chapter concludes the research and experimentation on mortgage arrears and has delivered an overview of the work completed as well as the results achieved. Possible areas of future research have been introduced whereby the experiment can be continued and improved upon.

The models produced as part of the experiment have been very successful at predicting arrears and have solved an acute business problem within Lender A. The experiment has shown that both customer spending habits and housing equity values can help to improve the predictive capability of a model to predict arrears. Further data to be added to the model could be identified from some of the future work outlined in Section 7.5.

Sampling methods that were applied to the dataset to address the imbalance in the classes have shown an increase in the accuracy of Recall, though at the cost of the accuracy of the Specificity and Precision rates. Synthetic sampling gave the highest Recall scores, though the other scores for these models decreased drastically. Both over and under sampling have achieved better Recall rates than the models with no sampling applied, and have returned relatively good scores for the other measures. Overall, sampling helped to achieve models with better scores, and it also reduced the associated misclassification costs of the models produced. Alternative methods could be utilised on the models to address the balance of the data in the dataset, including changing the parameters and thresholds though this would have to be explored in future work.

Both the Head of Data and the Head of Strategy in FSG were very impressed with the speed at which the models were built and deployed as well as the output produced. As a result of the work carried out on the model, FSG have requested a number of follow on items of work that will complement the model and further develop the work already completed. These pieces of analysis and modelling will give a greater understanding to Lender A of what is happening with the current arrears situation, and how best to address it going forward. Ultimately by managing the mortgage arrears within the bank better,

BIBLIOGRAPHY

Bellotti, T., Crook, J., 2012. Loss given default models incorporating macroeconomic variables for credit cards. Int. J. Forecast. 28, 171–182. doi:10.1016/j.ijforecast.2010.08.005

Bhutta, N., Shan, H., Dokko, J., 2010. The depth of negative equity and mortgage default decisions.

Blumer, A., Ehrenfeucht, A., Haussler, D., Warmuth, M.K., 1987. Occam's razor. Inf. Process. Lett. 24, 377–380.

Central Bank of Ireland, 2013. Residential Mortgage Arrears and Repossession Statistics: quarter 3 2013.

Central Statistics Office, 2008. Construction and Housing in Ireland.

Chairi, I., Alaoui, S., Lyhyaoui, A., 2012. Learning from imbalanced data using methods of sample selection, in: 2012 International Conference on Multimedia Computing and Systems (ICMCS). Presented at the 2012 International Conference on Multimedia Computing and Systems (ICMCS), pp. 254–257. doi:10.1109/ICMCS.2012.6320291

Chaudhuri, A., De, K., 2011. Fuzzy Support Vector Machine for bankruptcy prediction. Appl. Soft Comput. 11, 2472–2486. doi:10.1016/j.asoc.2010.10.003

Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P., 2011. SMOTE: synthetic minority over-sampling technique. ArXiv Prepr. ArXiv11061813.

Chen, M.-C., Chen, L.-S., Hsu, C.-C., Zeng, W.-R., 2008. An information granulation based data mining approach for classifying imbalanced data. Inf. Sci. 178, 3214–3227. doi:10.1016/j.ins.2008.03.018

Coenen, F., 2011. Data mining: past, present and future. Knowl. Eng. Rev. 26, 25–29.

Connor, G., Flavin, T., 2013. Irish Mortgage Default Optionality.

Connor, G., Flavin, T., O'Kelly, B., 2012. The US and Irish credit crises: Their distinctive differences and common features. J. Int. Money Finance 31, 60–79.

Delen, D., Walker, G., Kadam, A., 2005. Predicting breast cancer survivability: a comparison of three data mining methods. Artif. Intell. Med. 34, 113–127.

Domingos, P., 1999. The role of Occam's razor in knowledge discovery. Data Min. Knowl. Discov. 3, 409–425.

Duffy, D., ESRI, O'Hanlon, N., 2013. Negative equity in the Irish housing market: Estimates using loan level data.

Dufhues, T., Buchenrieder, G., Quoc, H.D., Munkung, N., 2011. Social capital and loan repayment performance in Southeast Asia. J. Socio-Econ. 40, 679–691. doi:10.1016/j.socec.2011.05.007

Ellis, L., 2008a. The housing meltdown: Why did it happen in the United States?

Ellis, L., 2008b. How many in negative equity? The role of mortgage contract characteristics. BIS Q. Rev. 81.

Ellis, L., 2012. Prudent Mortgage Lending Standards Help Ensure Financial Stability, in: Address to the Australian Mortgage Conference, Sydney.

Elul, R., Souleles, N.S., Chomsisengphet, S., Glennon, D., Hunt, R., 2010. What" triggers" mortgage default? Am. Econ. Rev. 100, 490–494.

Eurostat — Statistical Office of the European Communities,, 2007. Eurostat yearbook 2006-07.

Fayyad, U.M., Piatetsky-Shapiro, G., Smyth, P., 1996. Knowledge Discovery and Data Mining: Towards a Unifying Framework., in: KDD. pp. 82–88.

Finlay, S.M., 2008. Multiple classifier architectures and their application to credit risk assessment.

Foote, C.L., Gerardi, K., Goette, L., Willen, P.S., 2008a. Subprime facts: What (we think) we know about the subprime crisis and what we don't. Public policy Discussion Papers, Federal Reserve Bank of Boston.

Foote, C.L., Gerardi, K., Willen, P.S., 2008b. Negative equity and foreclosure: Theory and evidence. J. Urban Econ. 64, 234–245. doi:10.1016/j.jue.2008.07.006

Frawley, W.J., Piatetsky-Shapiro, G., Matheus, C.J., 1992. Knowledge discovery in databases: An overview. AI Mag. 13, 57.

Freund, Y., Schapire, R.E., 1995. A desicion-theoretic generalization of on-line learning and an application to boosting, in: Computational Learning Theory. Springer, pp. 23–37.

Friedman, J.H., Kohavi, R., Yun, Y., 1996. Lazy decision trees, in: AAAI/IAAI, Vol. 1. pp. 717–724.

Gathergood, J., 2012. Self-control, financial literacy and consumer over-indebtedness. J. Econ. Psychol. 33, 590–602. doi:10.1016/j.joep.2011.11.006

Gerardi, K., Goette, L., Meier, S., 2013. Numerical ability predicts mortgage default. Proc. Natl. Acad. Sci. 110, 11267–11271.

Gerardi, K., Shapiro, A.H., Willen, P.S., 2008. Subprime outcomes: Risky mortgages, homeownership experiences, and foreclosures. Working paper series//Federal Reserve Bank of Boston.

Ghent, A.C., Kudlyak, M., 2011. Recourse and residential mortgage default: Evidence from US states. Rev. Financ. Stud. 24, 3139–3186.

Goggin, J., Holton, S., Kelly, J., Lydon, R., McQuinn, K., 2012. Variable Mortgage Rate Pricing in Ireland. Econ. Lett. 2.

Guiso, L., Sapienza, P., Zingales, L., 2013. The determinants of attitudes towards strategic default on mortgages. J. Finance.

Gyourko, J., Tracy, J., 2014. Reconciling theory and empirics on the role of unemployment in mortgage default. J. Urban Econ. 80, 87–96. doi:10.1016/j.jue.2013.10.005

Han, H., Wang, W.-Y., Mao, B.-H., 2005. Borderline-SMOTE: A new over-sampling method in imbalanced data sets learning, in: Advances in Intelligent Computing. Springer, pp. 878–887.

Hassan, A.K.I., Abraham, A., 2013. Modeling consumer loan default prediction using neural netware, in: 2013 International Conference on Computing, Electrical and Electronics Engineering (ICCEEE). Presented at the 2013 International Conference on Computing, Electrical and Electronics Engineering (ICCEEE), pp. 239–243. doi:10.1109/ICCEEE.2013.6633940

Haykin, S., 1994. Neural networks: a comprehensive foundation. Prentice Hall PTR.

He, H., Bai, Y., Garcia, E.A., Li, S., 2008. ADASYN: Adaptive synthetic sampling approach for imbalanced learning, in: Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence). IEEE International Joint Conference on. IEEE, pp. 1322–1328.

He, H., Garcia, E.A., 2009. Learning from Imbalanced Data. IEEE Trans. Knowl. Data Eng. 21, 1263–1284. doi:10.1109/TKDE.2008.239

Heo, H., Park, H., Kim, N., Lee, J., 2009. Prediction of credit delinquents using locally transductive multi-layer perceptron. Neurocomputing 73, 169–175. doi:10.1016/j.neucom.2009.02.025

Honohan, P., 2009. Resolving Ireland's Banking Crisis. Econ. Soc. Rev. 40.

Iyer, R., Puri, M., 2010. Understanding bank runs: The importance of depositor-bank relationships and networks," forthcoming. Am. Econ. Rev.

Jagannatha Reddy, M.V., Kavitha, B., 2010. Neural Networks for Prediction of Loan Default Using Attribute Relevance Analysis, in: International Conference on Signal Acquisition and Processing, 2010. ICSAP '10. Presented at the International Conference on Signal Acquisition and Processing, 2010. ICSAP '10, pp. 274–277. doi:10.1109/ICSAP.2010.10

Kaastra, I., Boyd, M.S., 1995. Forecasting futures trading volume using neural networks. J. Futur. Mark. 15, 953–970.

Keese, M., 2009. Triggers and determinants of severe household indebtedness in Germany.

Kelly, M., 2009. The Irish credit bubble. University College Dublin, School of Economics.

Kelly, R., 2012. House Prices, Unemployment and Irish Mortgage Losses. Res. Tech. Pap. Cent. Bank Irel.

Klapper, L., Lusardi, A., Panos, G.A., 2013. Financial literacy and its consequences: Evidence from Russia during the financial crisis. J. Bank. Finance 37, 3904–3923. doi:10.1016/j.jbankfin.2013.07.014

Lane, P.R., 2011. The Irish Crisis.

Li, W., White, M.J., Zhu, N., 2010. Did Bankruptcy Reform Cause Mortgage Default to Rise? National Bureau of Economic Research.

Liu, X.-Y., Wu, J., Zhou, Z.-H., 2009. Exploratory undersampling for class-imbalance learning. Syst. Man Cybern. Part B Cybern. IEEE Trans. On 39, 539–550.

Liu, Y., Weisberg, R.H., Hu, C., Zheng, L., 2011. Tracking the Deepwater Horizon oil spill: A modeling perspective. Eos Trans. Am. Geophys. Union 92, 45–46.

McCarthy, Y., 2014. Dis-entangling the mortgage arrears crisis The role of the labour market, income volatility and negative equity.

McCarthy, Y., McQuinn, K., 2011. How Are Irish Households Coping with their Mortgage Repayments? Information from the Survey on Income and Living Conditions. Econ. Soc. Rev. 42.

McCarthy, Y., McQuinn, K., 2013. Credit conditions in a boom and bust property market.

McQuinn, K., O'Reilly, G., 2008. Assessing the role of income and interest rates in determining house prices. Econ. Model. 25, 377–390. doi:10.1016/j.econmod.2007.06.010

Neter, J., Kutner, M.H., Nachtsheim, C.J., Wasserman, W., 1996. Applied linear statistical models. Irwin Chicago.

Norris, M., Coates, D., House, M., 2010. How Housing Killed the Celtic Tiger: Anatomy, Consequences and Lessons of Ireland's Mortgage Boom and Bust, 2000-2009. Hous. Next 20.

O'Toole, F., 1993. Tax reform since the Commission on Taxation, in: A Research Seminar of the Foundation for Fiscal Studies.

Odeh, O., Koduru, P., Featherstone, A., Das, S., Welch, S.M., 2011. A multi-objective approach for the prediction of loan defaults. Expert Syst. Appl. 38, 8850–8857. doi:10.1016/j.eswa.2011.01.096

Oreski, S., Oreski, D., Oreski, G., 2012. Hybrid system with genetic algorithm and artificial neural networks and its application to retail credit risk assessment. Expert Syst. Appl. 39, 12605–12617. doi:10.1016/j.eswa.2012.05.023

Quinlan, J.R., 1986. Induction of decision trees. Mach. Learn. 1, 81–106.

RTÉ.ie Business news, 2013. Most recent changes in ECB's interest rate: [WWW Document]. RTE.ie. URL http://www.rte.ie/news/business/economy/2012/0327/315217-interestrates/ (accessed 2.25.14).

Russell, H., Maître, B., Donnelly, N., 2011. Financial Exclusion and Over-indebtedness in Irish Households. Department of Community, Equality & Gaeltacht Affairs and Economic and Social Research Institute.

Schapire, R.E., 2003. The boosting approach to machine learning: An overview, in: Nonlinear Estimation and Classification. Springer, pp. 149–171.

Scheurmann, E., Matthews, C., 2005. Neural network classifers in arrears management, in: Artificial Neural Networks: Formal Models and Their Applications–ICANN 2005. Springer, pp. 325–330.

Setiono, R., Baesens, B., Mues, C., 2008. Recursive Neural Network Rule Extraction for Data With Mixed Attributes. IEEE Trans. Neural Netw. 19, 299–307. doi:10.1109/TNN.2007.908641

Shin, K.-S., Lee, T.S., Kim, H., 2005. An application of support vector machines in bankruptcy prediction model. Expert Syst. Appl. 28, 127–135.

Silverman, N., Suchard, M.A., 2013. PREDICTING HORSE RACE WINNERS THROUGH REGULARIZED CONDITIONAL LOGISTIC REGRESSION WITH FRAILTY. J. Predict. Mark. 7.

Srinivasan, B.V., Gnanasambandam, N., Zhao, S., Minhas, R., 2011. Domain-Specific Adaptation of a Partial Least Squares Regression Model for Loan Defaults Prediction, in: 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW). Presented at the 2011 IEEE 11th International Conference on Data Mining Workshops (ICDMW), pp. 474–479. doi:10.1109/ICDMW.2011.69

Stekler, H.O., Klein, A., 2012. Predicting the outcomes of NCAA basketball championship games. J. Quant. Anal. Sports 8, 3.

Sun, J., Li, H., 2012. Financial distress prediction using support vector machines: Ensemble vs. individual. Appl. Soft Comput. 12, 2254–2265. doi:10.1016/j.asoc.2012.03.028

Trautmann, S.T., Vlahu, R., 2013. Strategic loan defaults and coordination: An experimental analysis. J. Bank. Finance 37, 747–760. doi:10.1016/j.jbankfin.2012.10.019

Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M., 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM 10, 178–185.

Utrilla, J.M., Constantinou, N., 2010. Could the subprime crisis have been predicted? A mortgage risk modeling approach.

Vapnik, V.N., 1999. An overview of statistical learning theory. Neural Netw. IEEE Trans. On 10, 988–999.

Wang, B., Liu, Y., Hao, Y., Liu, S., 2007. Defaults Assessment of Mortgage Loan with Rough Set and SVM, in: 2007 International Conference on Computational Intelligence and Security. Presented at the 2007 International Conference on Computational Intelligence and Security, pp. 981–985. doi:10.1109/CIS.2007.159

White, B.T., 2010. Underwater and not walking away: shame, fear and the social management of the housing crisis.

APPENDIX A – ALL DATA ITEMS USED IN THE MODEL

Column Name	Data Type	Description
HM_LENDING_APPL_NO	Integer	The mortgage identifier
ARREARS_IND	Integer2	Whether the mortgage has arrears or not - this is the target
APPL_YEAR	Varchar	The year of the application for the mortgage
MIN_OPEN_YEAR	Varchar	The first year one of the loans associated with the mortgage was opened
NUM_CUSTS	Varchar	The number of customers associated with the mortgage
MORTGAGE_PRICING	Varchar	The interest rate pricing of the mortgage i.e. fixed, variable etc.
NUM_ACS	Varchar	The number of loan accounts associated with the mortgage
CHANNEL	Varchar	The channel the application came through
CURR_MAX_CR_GRADE	Varchar	The maximum credit grade associated with the accounts on the mortgage
PRIOR_ARREARS	Varchar	Flag to indicate if there was ever prior arrears on the mortgage
ARREARS_IN_LAST_6_MTHS	Varchar	Flag to indicate if there was prior arrears on the mortgage in the last 6 months
ARREARS_IN_LAST_12_MTHS	Varchar	Flag to indicate if there was prior arrears on the mortgage in the last 12 months
MORT_TERM	Varchar	The term of the mortgage
NUM_YRS_OPEN	Varchar	The number of years the mortgage has been open
NUM_YRS_LEFT	Varchar	The number of years left on the mortgage
FIRST_TIME_BUYER_IND	Char	Flag to indicate if the borrower was a first time buyer at the time of application
APPL_CHANNEL_CDE	Char	The application channel code
EQUITY_PERCENT	Varchar	The percentage equity in the property based on estimated property value
CURRENT_LTV	Varchar	The current Loan to Value (LTV) rate
HOUSE_TYPE	Varchar	The type of property e.g. bungalow, detached etc.
NO_BEDROOMS_CNT	Varchar	Number of bedrooms in the property
YEAR_OF_CONSTRUCTION	Varchar	The year the property was built
COUNTY	Varchar	The county the property is located in (Postcode if in Dublin)
DEPENDENT_CHILDREN_IND	Char(1)	Flag to indicate the borrower has dependent children
CUST_MIN_AGE	Varchar	The minimum age of the borrowers associated with the mortgage
CUST_MAX_AGE	Varchar	The maximum age of the borrowers associated with the mortgage
CUST_AVG_AGE	Varchar	The average age of the borrowers associated with the mortgage.
MTH_6_EMTS_CHANGE	Decimal	The change in salary amount lodged to the account 6 months ago.
MTH_5_EMTS_CHANGE	Decimal	The change in salary amount lodged to the account 5 months ago.
MTH_4_EMTS_CHANGE	Decimal	The change in salary amount lodged to the account 4 months ago.
MTH_3_EMTS_CHANGE	Decimal	The change in salary amount lodged to the account 3 months ago.
MTH_2_EMTS_CHANGE	Decimal	The change in salary amount lodged to the account 2 months ago.
MTH_1_EMTS_CHANGE	Decimal	The change in salary amount lodged to the account 1 months ago.
MTH_6_CIF_CHANGE	Decimal	The change in salary amount recorded on the customer system 6 months ago.
MTH_5_CIF_CHANGE	Decimal	The change in salary amount recorded on the customer system 5 months ago.
MTH_4_CIF_CHANGE	Decimal	The change in salary amount recorded on the customer system 4 months ago.
MTH_3_CIF_CHANGE	Decimal	The change in salary amount recorded on the customer system 3 months ago.
MTH_2_CIF_CHANGE	Decimal	The change in salary amount recorded on the customer system 2 months ago.
MTH_1_CIF_CHANGE	Decimal	The change in salary amount recorded on the customer system 1 months ago.
MTH_6_UNPAIDS_CNT	Integer	The number of unpaids on all accounts associated with the mortgage 6 months ago
MTH_5_UNPAIDS_CNT	Integer	The number of unpaids on all accounts associated with the mortgage 5 months ago
MTH_4_UNPAIDS_CNT	Integer	The number of unpaids on all accounts associated with the mortgage 4 months ago
MTH_3_UNPAIDS_CNT	Integer	The number of unpaids on all accounts associated with the mortgage 3 months ago
MTH_2_UNPAIDS_CNT	Integer	The number of unpaids on all accounts associated with the mortgage 2 months ago
MTH_1_UNPAIDS_CNT	Integer	The number of unpaids on all accounts associated with the mortgage 1 month ago
TOTAL_UNPAIDS_LAST_6_MTHS	Integer	Total number of unpaid transactions in the last 6 months
UNPAIDS_IN_PAST_MTH_IND	Char(1)	Flag to indicate if there have been unpaids in the past month associated with the mortgage.
UNPAIDS_IN_PAST_6_MTHS_IND	Char(1)	Flag to indicate if there have been unpaids in the past 6 months associated with the mortgage.
MTH_6_UTLISED_PERCENT	Decimal	The credit card balance as a % of the overall limit for the previous month
MTH_5_UTLISED_PERCENT	Decimal	The credit card balance as a % of the overall limit for the previous month
MTH_4_UTLISED_PERCENT	Decimal	The credit card balance as a % of the overall limit for the previous month
MTH_3_UTLISED_PERCENT	Decimal	The credit card balance as a % of the overall limit for the previous month
MTH_2_UTLISED_PERCENT	Decimal	The credit card balance as a % of the overall limit for the previous month
MTH_1_UTLISED_PERCENT	Decimal	The credit card balance as a % of the overall limit for the previous month
MTH_6_OVERLIMIT_PERCENT	Decimal	The overlimit amount for the credit cards as a % of the total card limit for the previous month

MTH_5_OVERLIMIT_PERCENT	Decimal	The overlimit amount for the credit cards as a % of the total card limit for the previous month
MTH_4_OVERLIMIT_PERCENT	Decimal	The overlimit amount for the credit cards as a % of the total card limit for the previous month
MTH_3_OVERLIMIT_PERCENT	Decimal	The overlimit amount for the credit cards as a % of the total card limit for the previous month
MTH_2_OVERLIMIT_PERCENT	Decimal	The overlimit amount for the credit cards as a % of the total card limit for the previous month
MTH_1_OVERLIMIT_PERCENT	Decimal	The overlimit amount for the credit cards as a % of the total card limit for the previous month
MTH_6_CC_CHANGE	Decimal	The change in the credit card utilisation from the previous month
MTH_5_CC_CHANGE	Decimal	The change in the credit card utilisation from the previous month
MTH_4_CC_CHANGE	Decimal	The change in the credit card utilisation from the previous month
MTH_3_CC_CHANGE	Decimal	The change in the credit card utilisation from the previous month
MTH_2_CC_CHANGE	Decimal	The change in the credit card utilisation from the previous month
MTH_1_CC_CHANGE	Decimal	The change in the credit card utilisation from the previous month
MTH_6_SAVINGS_CHANGE	Decimal	The % change in the total savings balance from the previous month
MTH_5_SAVINGS_CHANGE	Decimal	The % change in the total savings balance from the previous month
MTH_4_SAVINGS_CHANGE	Decimal	The % change in the total savings balance from the previous month
MTH_3_SAVINGS_CHANGE	Decimal	The % change in the total savings balance from the previous month
MTH_2_SAVINGS_CHANGE	Decimal	The % change in the total savings balance from the previous month
MTH_1_SAVINGS_CHANGE	Decimal	The % change in the total savings balance from the previous month
MTH_6_SPEND_PERCENT	Decimal	The Spend txns as % of all txns for the previous month
MTH_5_SPEND_PERCENT	Decimal	The Spend txns as % of all txns for the previous month
MTH_4_SPEND_PERCENT	Decimal	The Spend txns as % of all txns for the previous month
MTH_3_SPEND_PERCENT	Decimal	The Spend txns as % of all txns for the previous month
MTH_2_SPEND_PERCENT	Decimal	The Spend txns as % of all txns for the previous month
MTH_1_SPEND_PERCENT	Decimal	The Spend txns as % of all txns for the previous month
MTH_6_SPEND_CHANGE	Decimal	The Spend txn % change from the previous month
MTH_5_SPEND_CHANGE	Decimal	The Spend txn % change from the previous month
MTH_4_SPEND_CHANGE	Decimal	The Spend txn % change from the previous month
MTH_3_SPEND_CHANGE	Decimal	The Spend txn % change from the previous month
MTH_2_SPEND_CHANGE	Decimal	The Spend txn % change from the previous month
MTH_1_SPEND_CHANGE	Decimal	The Spend txn % change from the previous month
MTH_6_SAVE_PERCENT	Decimal	The Save txns as % of all txns for the previous month
MTH_5_SAVE_PERCENT	Decimal	The Save txns as % of all txns for the previous month
MTH_4_SAVE_PERCENT	Decimal	The Save txns as % of all txns for the previous month
MTH_3_SAVE_PERCENT	Decimal	The Save txns as % of all txns for the previous month
MTH_2_SAVE_PERCENT	Decimal	The Save txns as % of all txns for the previous month
MTH_1_SAVE_PERCENT	Decimal	The Save txns as % of all txns for the previous month
MTH_6_SAVE_CHANGE	Decimal	The Save txn % change from the previous month
MTH_5_SAVE_CHANGE	Decimal	The Save txn % change from the previous month
MTH_4_SAVE_CHANGE	Decimal	The Save txn % change from the previous month
MTH_3_SAVE_CHANGE	Decimal	The Save txn % change from the previous month
MTH_2_SAVE_CHANGE	Decimal	The Save txn % change from the previous month
MTH_1_SAVE_CHANGE	Decimal	The Save txn % change from the previous month
MTH_6_LIVE_PERCENT	Decimal	The Live txns as % of all txns for the previous month
MTH_5_LIVE_PERCENT	Decimal	The Live txns as % of all txns for the previous month
MTH_4_LIVE_PERCENT	Decimal	The Live txns as % of all txns for the previous month
MTH_3_LIVE_PERCENT	Decimal	The Live txns as % of all txns for the previous month
MTH_2_LIVE_PERCENT	Decimal	The Live txns as % of all txns for the previous month
MTH_1_LIVE_PERCENT	Decimal	The Live txns as % of all txns for the previous month
MTH_6_LIVE_CHANGE	Decimal	The Live txn % change from the previous month
MTH_5_LIVE_CHANGE	Decimal	The Live txn % change from the previous month
MTH_4_LIVE_CHANGE	Decimal	The Live txn % change from the previous month
MTH_3_LIVE_CHANGE	Decimal	The Live txn % change from the previous month
MTH_2_LIVE_CHANGE	Decimal	The Live txn % change from the previous month
MTH_1_LIVE_CHANGE	Decimal	The Live txn % change from the previous month
MTH_6_MORTGAGE_PERCENT	Decimal	The Mortgage txns as % of all txns for the previous month
MTH_5_MORTGAGE_PERCENT	Decimal	The Mortgage txns as % of all txns for the previous month
MTH_4_MORTGAGE_PERCENT	Decimal	The Mortgage txns as % of all txns for the previous month
MTH_3_MORTGAGE_PERCENT	Decimal	The Mortgage txns as % of all txns for the previous month

MTH_2_MORTGAGE_PERCENT	Decimal	The Mortgage txns as % of all txns for the previous month
MTH_1_MORTGAGE_PERCENT	Decimal	The Mortgage txns as % of all txns for the previous month
MTH_6_MORTGAGE_CHANGE	Decimal	The Mortgage txn % change from the previous month
MTH_5_MORTGAGE_CHANGE	Decimal	The Mortgage txn % change from the previous month
MTH_4_MORTGAGE_CHANGE	Decimal	The Mortgage txn % change from the previous month
MTH_3_MORTGAGE_CHANGE	Decimal	The Mortgage txn % change from the previous month
MTH_2_MORTGAGE_CHANGE	Decimal	The Mortgage txn % change from the previous month
MTH_1_MORTGAGE_CHANGE	Decimal	The Mortgage txn % change from the previous month
PC1	Decimal	Principal Component 1
PC2	Decimal	Principal Component 2
PC3	Decimal	Principal Component 3
PC4	Decimal	Principal Component 4
PC5	Decimal	Principal Component 5
PC6	Decimal	Principal Component 6
PC7	Decimal	Principal Component 7
PC8	Decimal	Principal Component 8
PC9	Decimal	Principal Component 9
PC10	Decimal	Principal Component 10
PC11	Decimal	Principal Component 11
PC12	Decimal	Principal Component 12
PC13	Decimal	Principal Component 13
PC14	Decimal	Principal Component 14
PC15	Decimal	Principal Component 15
PC16	Decimal	Principal Component 16
PC17	Decimal	Principal Component 17
PC18	Decimal	Principal Component 18
PC19	Decimal	Principal Component 19
PC20	Decimal	Principal Component 20
PC21	Decimal	Principal Component 21
PC22	Decimal	Principal Component 22
PC23	Decimal	Principal Component 23
PC24	Decimal	Principal Component 24
PC25	Decimal	Principal Component 25
PC26	Decimal	Principal Component 26
PC27	Decimal	Principal Component 27
PC28	Decimal	Principal Component 28
PC29	Decimal	Principal Component 29
PC30	Decimal	Principal Component 30
PC31	Decimal	Principal Component 31
PC32	Decimal	Principal Component 32
PC33	Decimal	Principal Component 33
PC34	Decimal	Principal Component 34
PC35	Decimal	Principal Component 35