2018

Clicking into Mortgage Arrears: A Study into Arrears Prediction with Clickstream Data

Gavin O'Brien
Technological University Dublin

Clicking into Mortgage Arrears: A Study into Arrears Prediction with Clickstream Data

Gavin O'Brien

D14128195

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

2018

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: ***Gavin O'Brien*** Gavin O'Brien

Date: **13th June 2018**

ABSTRACT

This research project investigates the predictive capability of clickstream data when used for the purpose of mortgage arrears prediction. With an ever growing number of people switching to digital channels to handle their daily banking requirements, there is a wealth of ever increasing online usage data, otherwise known as clickstream data. If leveraged correctly, this clickstream data can be a powerful data source for organisations as it provides detailed information about how their customers are interacting with their digital channels. Much of the current literature associated with clickstream data relates to organisations employing it within their customer relationship management mechanisms to build better relationships with their customers. There has been little investigation into the use of clickstream data in credit scoring or arrears prediction.

Since the financial meltdown of 2008, financial institutions have being obliged to have mechanisms in place to deal with mortgage accounts which are in arrears or have a risk of entering arrears. A potentially crucial step in this process is the ability of an institution to accurately predict which of their mortgage accounts may enter arrears. In addition to traditional demographical and transactional data, this research determines the impact clickstream data can have on an arrears prediction model. A multitude of binary classifiers were reviewed in this arrears prediction problem. Of these classifiers, ensembles models proved to be the highest performing models achieving reasonably high recall accuracies without the inclusion of clickstream data. Once clickstream data was added to the models, it led to marginal increases in accuracy, which was a positive result.

Key words: *mortgage arrears, clickstream, banking, classification, data mining, risk modelling*

ACKNOWLEDGEMENTS

I would like to express my sincere thanks to my supervisor, Seán O'Leary, for his assistance and invaluable guidance throughout this dissertation.

I would also like to thank the senior management of Lender A for giving me the opportunity to complete this MSc and also supporting me throughout this dissertation. I'd also like to thank them for allowing the use of Lender A's data. In addition to this, I would like to express sincere thanks to Kevin, Tommy, Eanna, Lorna, Peter, Yuri, Kulbhushan, Dave T, Aodhán and various other staff members from within Lender A for supporting me and allowing me time to work on this dissertation.

Finally I would like to thank my family, especially my Father, for being extremely supportive over the time I've being working on this dissertation and the entire MSc as without this support, I would not have been able to complete it.

This dissertation is dedicated to the memory of my mother, Noeleen.

TABLE OF CONTENTS

TABLE OF FIGURES

TABLE OF TABLES

1 INTRODUCTION

1.1 Overview and Background

This research will be conducted using data from a large Irish financial institution which operates throughout Ireland via both a traditional branch network and through various digital channels. To protect their anonymity, their name will not be used, and the pseudonym Lender A will be used instead.

During the 1980's, the Irish economy was bleak. Unemployment was averaging approximately 14% and over 40% of Irish tax payers were paying a marginal tax rate of 45% (Whelan, 2009). However this time came to an end in late 1990's when the Irish economy began to boom. Known as the Celtic Tiger, this prosperous time in Ireland was driven mainly by the construction trade and Ireland joining the single currency of the EU in 1999. During this time the construction boom led to an employment boom and consequently employment was nearing full levels. Due to this, the tax income for the government increased exponentially and government spending increased substantially. As more people were entering high salary employment, many wanted to get their foot onto the property ladder. This coupled with the desire of many people to upgrade their home led to housing demand outstripping supply which consequently led to house prices skyrocketing with the average price of a home in Dublin hitting €400,000 and €250,000 outside Dublin (Norris & Brooke, 2011). Prior to 1997, Irish financial institutions employed their own deposits to finance loans. After this, they began using interbank loans and international bonds to fund lending. However these source of funds came an abrupt halt in 2008 after the collapse Lehmann Brothers in late 2007. This collapse is attributed to Lehmann Brothers lending too much capital to the sub-prime mortgage market.

This collapse of the Irish mortgage continued until 2011 at which point only 12,834 mortgages were approved across all of the Irish financial institutions. However since 2011, the market has begun slowly to rise again as is show in figure 1.1. Consequently financial institutions are cautious with regards who they are lending to and also their current mortgage stock. After the financial meltdown of 2008, the Irish government introduced legislation in an attempt to prevent such substantial damage to the Irish economy occurring again due to the mortgage market collapse. Contained within this legislation were various measures aimed at mitigating the mortgage market becoming over-inflated which could lead to another collapse. One of these measures was the

creation of the Mortgage Arrears Resolution Process (MARP). Financial institutions must have mechanisms in place to cater for MARP which requires them to help borrows in arrears or at the risk of entering arrears.

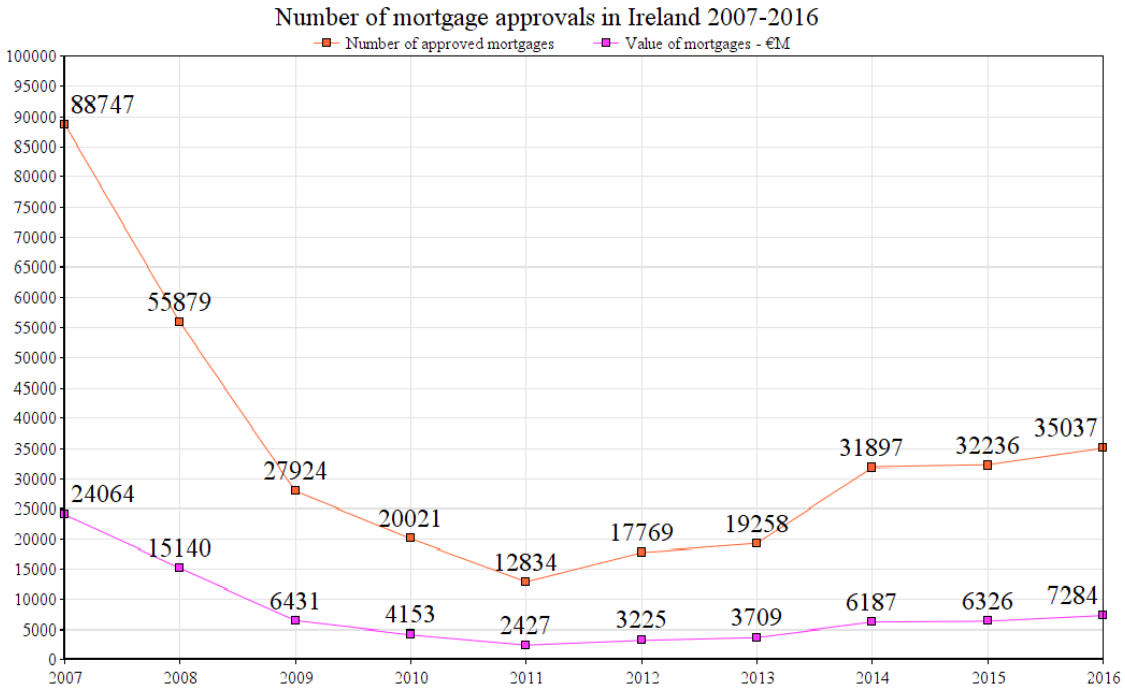

Figure 1.1 - Collapse and rise of Irish mortgage market 2007-2016; Data: (Department of Housing, Planning and Local Government, 2018)

Since the collapse of the Irish economy in 2007/2008, Irish financial institutions have been required to streamline their processes in an attempt to reduce costs. One driving factor of this streamlining is encouraging the adoption of digital channels which reduces the overall number of branches and branch staff required by financial institutions. A prime example of this framework is *N26*, a German based mobile bank which operates zero branches in Ireland but offers services similar to other financial institutions which have an entire network of branches. Additionally there are also various other Fintech organisations such as PayPal and Stripe which traditional institutions have to compete with in addition to their current competition in the Irish market.

Consequently institutions across Ireland and Europe are being forced to innovate and change. They have to employ technology and data in an attempt to become a customer centric organisation. This is the primary driver behind these institutions pushing their digital channels to their customers. Figure 1.2 shows the adoption rate of internet banking in the UK financial sector. This shows there is strong growth in the adoption of digital channels in the UK. Similarly 58% of Irish citizens used internet banking in 2017 (Taylor, 2018).

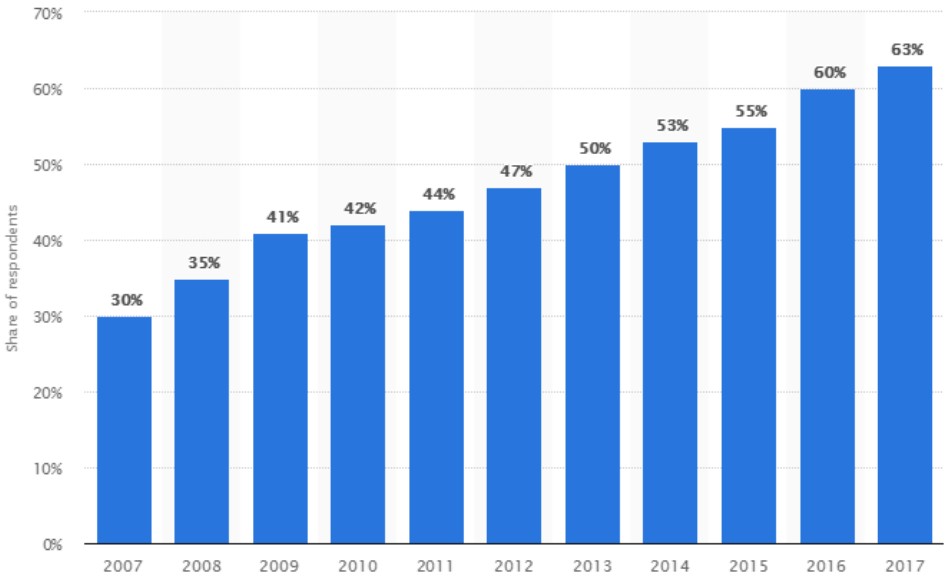

Figure 1.2 - Adoption of internet bank in the UK (Statista, 2018)

A by-product of this digital adoption is the institutions ability to monitor their customer interactions with their digital channels. This is the source of their clickstream data. Up until now, Lender A have only employed its clickstream data for driving direct sales to customers. For example if a customer used the loan calculator but did not follow through with applying for the loan, the customer may receive a phone call from direct banking enquiring whether they would like to apply for the loan or whether they had any queries with the application process.

Apart from this purpose, Lender A are driving little value from their clickstream data. This was the driving force behind this research. As Lender A must be compliant with Irish legislation, they are required to have a team which maintains their MARP. Within Lender A, the Arrears Support Unit (ASU) have this responsibility. If Lender A were able to predict accurately who may enter arrears, they may be able to engage with the customers pre-emptively and avoid further issues such as the potential of a customer default on the mortgage account.

Currently Lender A do not employ any model to predict if a mortgage account may enter arrears. They do have models to predict accounts which have a risk of default, which is an account in arrears over ninety days. Therefore if a model could be created which could predict if an account may enter arrears it should prove to be beneficial to Lender A. As the adoption of digital channels continue to grow in the financial sector, clickstream data should become more abundant for customers and therefore may provide

valuable insights to the financial life of a customer. A change in how a customer is interacting with Lender A's digital channels could indicate a change in their financial stability and therefore may potentially indicate that they may be about to enter arrears. For example, if a customer was logging on various times throughout the night, this could indicate they are under financial stress which could be a trigger for an arrears event to occur. Therefore close attention will be payed to the times a customer logs onto their online banking. In addition to utilising the times a customer connects with Lender A's digital channels, other features such as specific events they complete online may indicate potential problems in their personal funds. An example of this could be applying for a loan top up.

All of these clickstream features will be coupled with traditional features such as a customer's age, spending history and savings amount etc. in an attempt to create an accurate model which can predict if an account has the potential of entering arrears one month in advance of the arrears event occurring. This timeframe gives Lender A's ASU team suitable time to contact the customers and attempt to provide help to the customers by methods such as restructuring their mortgage.

1.2 Research Problem

Lender A's Enterprise Data Warehouse (EDW) contains a multitude of "traditional" data which could be employed as the input to a mortgage arrears prediction model. This research project aims to establish the impact clickstream data can have to the prediction capability of a mortgage arrears prediction model.

1.3 Research Objectives

The primary goal of this research is to determine if the addition of clickstream data can impact the performance of a mortgage arrears predication model. This research will be conducted utilising data from Lender A's EDW.

The research objectives are as follows:

- Study the up-to-date literature regarding the Irish mortgage market and clickstream data analysis.
- Examine and review research in the area of data mining to understand modern approaches to two-class classification problems.
- Aggregate various data sources to a singular record for a mortgage account to allow for the creation of a single analytical base table (ABT) which will contain

all of the features in addition to an ARREARS_IND flag indicating the arrears status of the account.

- Design experiments to test the hypothesis
- Implement various predictive modelling algorithms on a base dataset containing only traditional features to understand how the models are performing and also to establish the highest performing models.
- Using the highest performing models from the point above, add clickstream data to the source dataset and run the models once again to obtain the model performance with clickstream data.
- Preform an investigative analysis using a dataset containing only clickstream features to establish the power of clickstream features for mortgage arrears prediction.
- Analyse the results of the experiments to establish the impact clickstream data had on the models. Also attempt to understand classification errors and poorly performing models.
- Evaluate the success or failure of the experiments.
- Define how this research could be expanded upon and what future work could be carried out in this area.

1.4 Research Methodology

The empirical research untaken for this dissertation will involve performing experimentation on a mortgage accounts historical transactional, demographical and clickstream data. Suitable evaluation techniques will be employed to determine the success of these experiments. All models created and implemented during the experimentation will be measured using statistical methods to allow for the comparison of their performance. The area of the research is essentially a two class classification problem. In addition to this, other areas are mortgage arrears classification and clickstream data analysis.

Two class classification is a core pillar of Data Mining (DM). DM is the process of attempting to predict future events based on historical events. This challenge of predicting future events from historical events is not a new challenge. In 1992, (Frawley, Piatetsky-Shapiro, & Matheus, 1992) proposed the framework of *Knowledge Discovery in Databases (KDD).* They defined KDD as "*The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data.*" (Frawley,

Piatetsky-Shapiro, & Matheus, 1992). This framework has steps to enable a researcher to go from raw data through to deriving useable insights from the data.

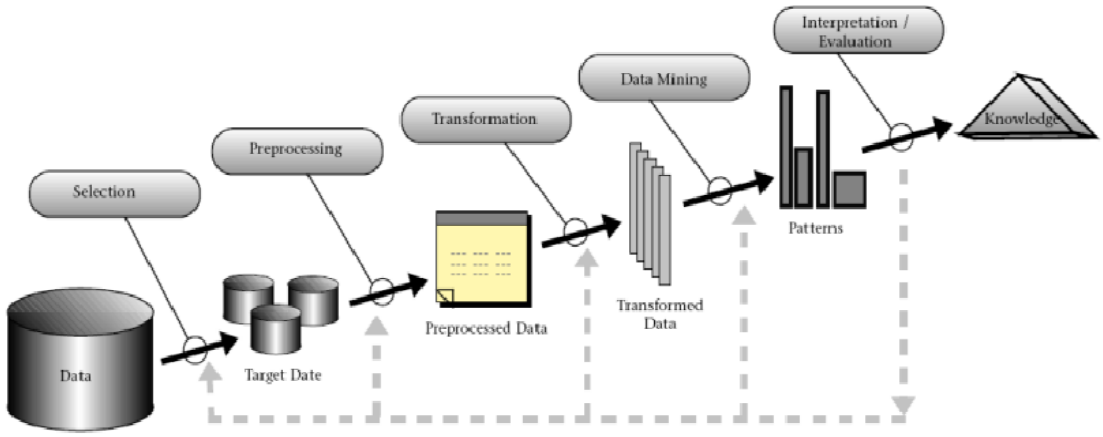

Figure 1.3 - KDD Process (Frawley, Piatetsky-Shapiro, & Matheus, 1992)

In addition to KDD, the Cross Industry Standard Process – Data Mining (CRISP-DM) will be leveraged to perform the experiments. The CRISP-DM methodology is a data mining process approach which defines commonly used techniques for the entire roadmap of a data mining project. This process involved six steps and will be discussed further in subsection *3.2* and can be seen in figure 1.4.

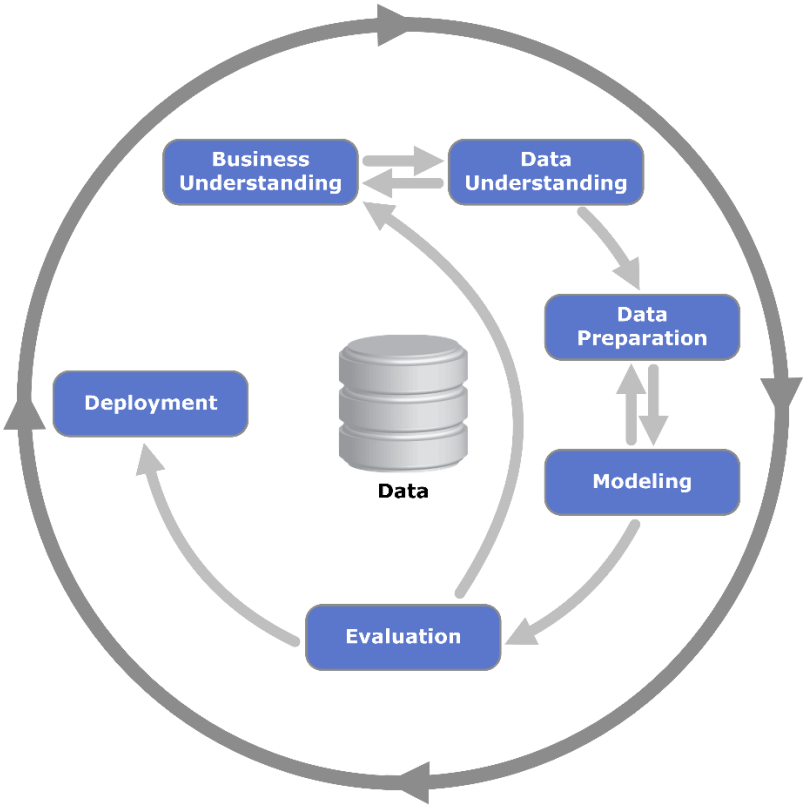

Figure 1.4 - CRISP-DM Methodology

1.5 Scope and Limitations

The project will focus on attempting to predict whether a mortgage account has a risk of entering arrears one month in advance of the arrear event occurring. Principle dwelling home (PDH) mortgages will be the focus of this research while buy-to-let (BTL) mortgages are excluded. BTL mortgages are excluded as the customer who is paying the mortgage would very rarely own one property and also generally have bought the property as an investment not a home. Consequently the entire PDH mortgage stock of Lender A will be utilised for the research.

The previous six months of data will be examined up to the date of prediction or the date of arrears for an account. For example, if the model is attempting to predict the account arrear status as of Sept 2017, the period which will be examined is the previous six months, Mar 2017 – Aug 2017. For all accounts not in arrears, this is the date range which will be used. If an account is in arrears, the date range will be the six months previous to the arrears event occurring. All demographical, transactional and clickstream data will be acquired from this date range and be aggregated to one record which will be added to an ABT. This ABT will contain one record per mortgage account and will be the source dataset for all modelling purposes.

All of the data used for the experiments will be extracted from Lender A's EDW. All of the "traditional" features such as a person's age, employment status and transactional history etc. are captured on numerous source system and fed into the EDW via various methods with little transformation (I.E. the EDW is an historical view of the source systems). The clickstream data is treated slightly differently. It is originally captured in an Hadoop environment via a MapReduce process and loaded into the EDW via and ETL tool. During this loading process, the clickstream/digital events are created as per rules pre-defined by Lender A. These are the rules which allow for the digital events to be categorised. This event classification will be leveraged heavily for the purposes of this research as certain events will be used as input features to modelling algorithms.

1.6 Document Outline

The remainder of this dissertation is organised into the following chapters; the Literature Review of the project area and data mining, Design of the experiments and Methodology of the research, Implementation and Results of the experiments and finally, the Conclusion.

Chapter 2 presents an encompassing view of the Irish mortgage market dating from pre Celtic Tiger Ireland through to a modern day mortgage market. In addition to this, clickstream data are presented along with their various uses and capture methods. Finally data mining techniques are examined including model evaluation techniques and methods for handling skewed datasets.

Chapter 3 discusses the methodology and design of the experiments carried out to determine the impact clickstream data can have on a mortgage arrears prediction model. The various different data sources are presented along with the software employed to implement the experiments. Finally the evaluation techniques used to determine the performance of the models are discussed.

Chapter 4 details the implementation of the experiments along with presenting the results. Detail of certain paths of research are discussed along with the rationale for choosing the paths based on the outcome of the experiments. Finally the results of all of the experiments are discussed and evaluated.

Chapter 5 finishes by reviewing the outcome of the experiments undertaken to establish the impact clickstream data had on a mortgage arrears predication model and concludes with further research and experiments which could be done in this field.

2 LITERATURE REVIEW

2.1 Introduction

This chapter will cover the relevant literature in relation to predictive modelling, mortgages and clickstream data, paying close attention to the Irish mortgage market and the banking industry in general. The insights gained from the literature review will help to determine the direction of this research. An in depth look into the Irish mortgage market both during and post Celtic Tiger Ireland will be presented followed by a review of clickstream data with a discussion of how it is captured and how it can be used. The key methodologies employed in the construction of predictive models will also be examined and accompanied with basic examples of how they function. In addition to this, various methods of evaluating the performance of predictive models will also be examined and justification for choosing certain evaluation methods over other evaluation methods will be presented.

As with many other predictive modelling issues, mortgage arrears fall into the category of skewed datasets. That is, mortgages in arrears are often a far smaller proportion of the dataset compared to mortgages which aren't in arrears. This can lead to difficulties in the modelling process and consequently the current literature on working with skewed datasets will also be reviewed.

2.2 Mortgages

In this section, there will be an in-depth examination of the Irish mortgage market, mortgage arrears and how people enter into arrears and how financial institutions deal with mortgage arrears. As the focus of this research is a mortgage lender in the Irish financial market, the main focus of this section will be the Irish mortgage market from before the Celtic Tiger through to the economic collapse of 2007/2008 and up until the modern day market.

2.2.1 Irish Mortgage Market during the Celtic Tiger

Between the late 1990's and late 2007, the Irish economy was booming. Known as the "Celtic Tiger", this prosperous time in Irish history contained exponential growth and led to an explosion in the demand for housing. However this was an entirely new situation to Ireland. The previous decade was bleak with high unemployment rates averaging approximately 14% and over 40% of Irish taxpayers paying a marginal rate of 45% in tax (Whelan, 2009). One of the main driving forces behind this prosperous period was Ireland joining the single currency of the EU in 1999 with another being the

construction trade. During this period, compared to the rest of the EU with the exception of Spain, Ireland was by far the largest constructor of housing, with 18 units per 1000 of the population being completed to an average of 5.3 units per 1000 of the population for the rest of the EU (CSO, 2008). This construction boom led to an employment boom which in-turn led to salaries in all sector rising to uncompetitive levels which consequently led to substantial increase in tax revenues which led to unprecedented levels of governmental spending. A result of the employment boom was the level of employment nearly doubling from 1990 to 2007, as can be seen in figure 2.1.

Figure 2.1 - Employment in Ireland from 1990 – 2010 (Lane P. , 2011)

As employment was near full levels, more and more people were attempting to get their foot onto the property ladder. This coupled with the wish of many people to upgrade their housing, led to the demand for housing outstripping supply which in turn led to explosive increases in the pricing of houses (Flavin & Connor, 2013). Figure 2.2 shows the increase of house prices across the country from 1996 through to 2007, before collapsing when the global financial crisis began to bite.

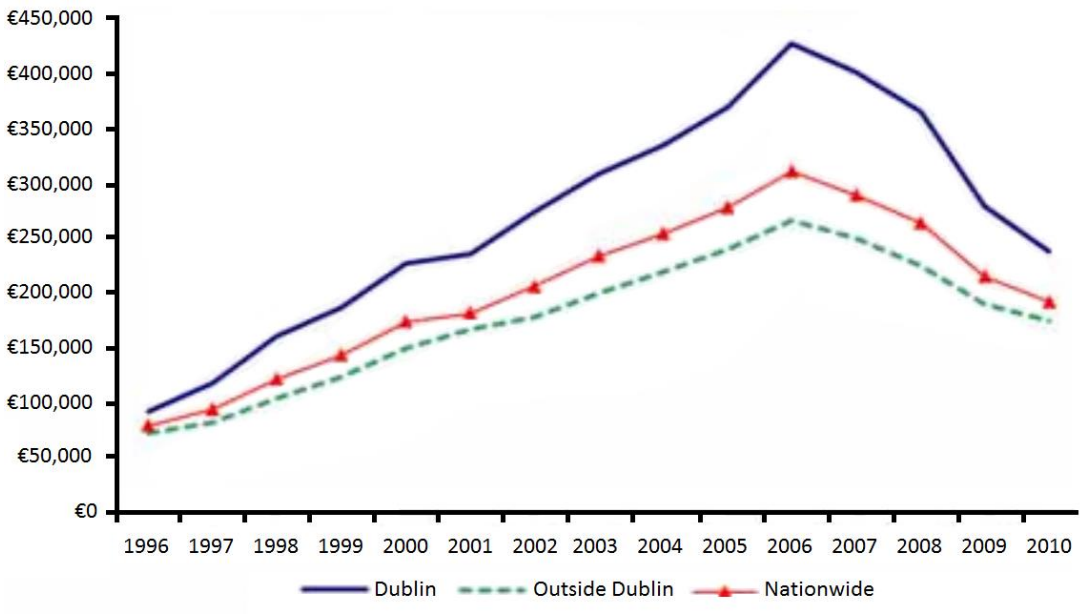

Figure 2.2 - House price increases in Ireland form 1996 – 2010 (Norris & Brooke, 2011)

As can be seen in figure 2.3, the rise of salaries did not keep up with the rise in house prices. In 1995, it took approximately four years earning to purchase a house across the country, where-as in 2007 it took seventeen years of earning to purchase a second-hand home in Dublin, and an average of 12 years across the country.

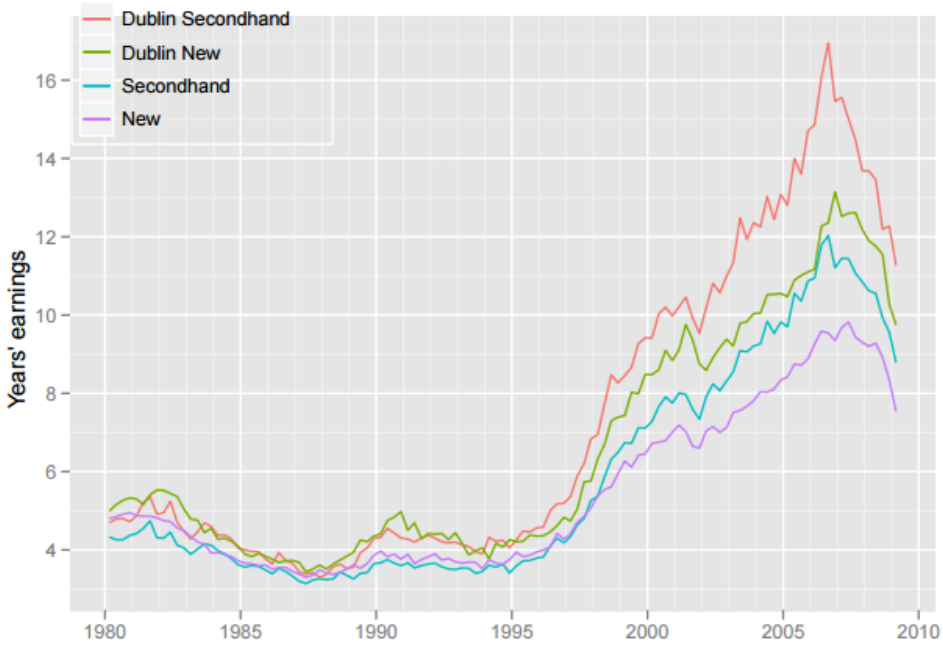

Figure 2.3 – Average number of years' salary to purchase a home from 1980 – 2010 (Kelly M. , 2009)

This all eventually accumulated in 2009 to where the size of the mortgage market grew to over €150B, which was over 100% of the late 2009 Irish GNP. Of this €150B, only around €40B was secured by assets (Kelly, 2009).

In the years leading up to the "Celtic Tiger", Irish financial institutions employed their customer's deposits as their source of funds for their lending. However this changed in 1997 when the institutions began using interbank loans and international bonds as a funding source for their lending. As of 2008, this led to over 50% of source of funds for loans coming from funds for which the institutions did not have deposits (Kelly, 2009). Consequently, the growth of the Irish mortgage market increased rapidly. The size of the market in 1997 was approximately €20B (in 2009 prices), with half of this being advanced to property developers. By 2008, the market had grown to €140 Billion, with €110B advanced to property developers (Roche, 2014). After the collapse of the Lehmann Brothers in 2007, the interbank loans and international bonds on which the Irish banks had become entirely dependent dried up and consequently the institutions were on the verge of collapse which would have been catastrophic for Ireland. To prevent this occurring, in an unprecedented move the Irish government stepped in and implemented a blanket guarantee scheme on all deposits at each of the covered institutions. In Europe, Denmark was the only other nation to impose a blanket guarantee on their financial institutions (Kelly, 2009).

2.2.2 Irish Mortgage Market post Celtic Tiger

The proposal by the Irish government to impose the blanket guarantee on Irish financial institutions was decided at a late night meeting with the financial institutions involved. While it has never been publicised, the most likely justification is that the government still believed the liquidity problems of the Irish institutions merely mirrored the instability of the global financial market in the wake of the Lehmann Brothers collapse (Kelly, 2009). The institutions shares prices began to steadily decline around March 2007, and despite the government's efforts, multiple financial institutions continued to decline which led to the government nationalising various institutions such as Anglo Irish Bank and also announcing the creation of a "bad bank", which was called the National Assets Management Agency (NAMA).

The purpose of NAMA was simple. It was to purchase non-preforming loans from Irish financial institutions in an attempt to recapitalise them. For example, if *Institution A* had €20B in capital and €150B in loans with €40B of developmental loans, and lost 50% of these developmental loans, their capital is wiped out and therefore they become insolvent. If NAMA was to purchase the €40B of development loans, the institutions becomes recapitalised and therefore can continue to operate. However a substantial

downfall of this is the Irish taxpayer faces a significant loss while the institutions shareholders face little to no repercussions (Whelan, 2009). Generally, NAMA is considered an international success story for dealing with bad assets (Schoenmaker, 2015).

However NAMA could not provide assistance on loans less than €20 million or to mortgages holders struggling to meet their financial commitments. Consequently this led to many mortgage holders turning to the Money Advice and Budgeting Services (MABS). MABS is the Irish government's money advice service offering free assistance and information about options available to borrowers in difficulty to allow them to meet their financial commitments. They also offer to act as an intermediary between the mortgage holders and the financial institutions to ensure the borrowers legal rights are upheld (Norris & Brooke, 2011). Figure 2.4 shows the number of clients turning to MABS for assistance grew steadily from 2008 through to 2010.

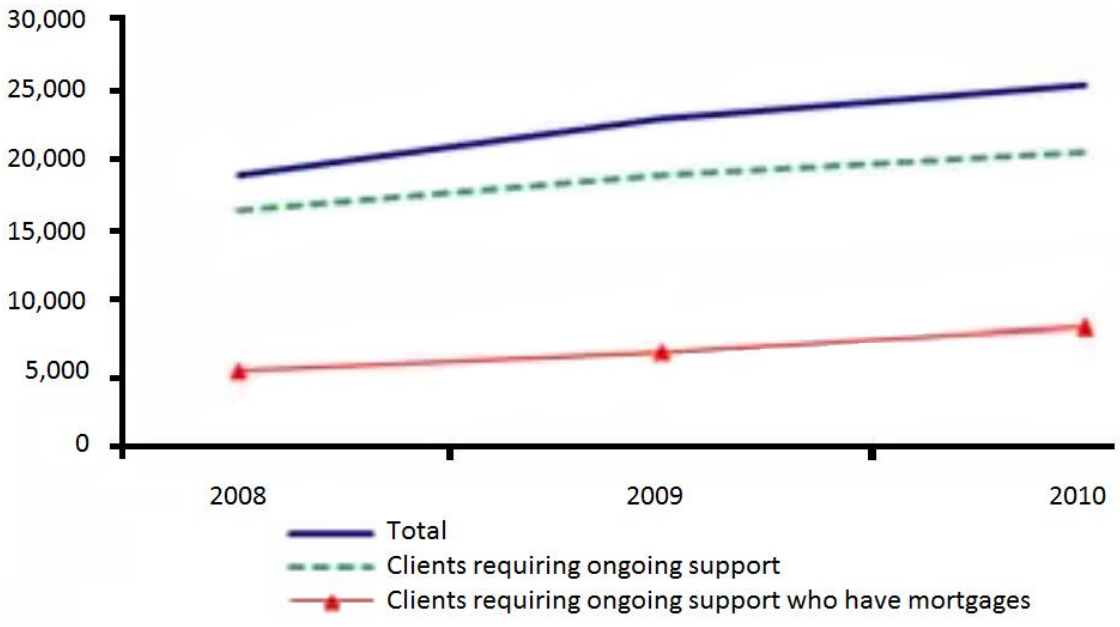

Figure 2.4 - Number of MABS clients from 2008 – 2010 (Norris & Brooke, 2011)

Between 2007 and 2012, the average price of homes in Ireland tumbled by 50% while at the same time unemployment rose from 5 to 15% (Kelly & O'Malley, 2016). Dublin was hit hardest by this collapse in house pricing where the average drop was 55% while the rest of the country fared only slightly better with a drop of 43%. As a result of this, there was a multitude of mortgages with high repayment burden and many which had entered negative equity. Negative equity occurs when the value of the home, on which the mortgage is secured, drops below the amount owed to repay the mortgage and

consequently if the home is sold the mortgage lender cannot be fully repaid without the mortgage holder finding a further source of funds (EBS, 2017).

As mentioned in subsection *2.2.3*, a mortgage arrears event is deemed to have occurred when a borrower has failed to make a mortgage repayment by a defined due date. The number of mortgages in arrears has continued to decline steadily each quarter since early 2014. As of Q4 2017, the number of mortgages in arrears in Ireland was a total number of 70,448, which is approx. 10% of the PDH mortgage market (Central Bank of Ireland, 2018). Further to this, 118,477 mortgages were classified as restructured with 87% of these mortgages deemed as meeting their restructuring obligations, which is a slight decrease from the Q3 2017. This is the eighteenth consecutive quarter where the number of mortgages in arrears have been declining since 2014 (Central Bank of Ireland, 2018).

In response to the collapse of the mortgage market and the end of the Celtic Tiger, a number of governmental policies changes were implemented in an attempt to reduce future risk. These changes included an introduction of a Mortgage Arrears Resolution Process (MARP) which all financial institutions must comply with in an attempt to help borrowers who have either entered arrears or may be of risk entering arrears. Consequently, if this research can be used to determine who may be of risk of entering arrears, it would be beneficial to the teams from Lender A who are involved in the MARP. Along with the introduction of MARP, there was also a moratorium on the repossessions of primary residence for 12 months since the initial arrear occurring (Norris & Brooke, 2011).

In addition to this, there were also limits brought into force to alleviate the possibility of borrowers receiving more than they can afford. This includes a limit of 3.5 Loan-to-income (LTI) on mortgages (that is, if a person's salary is €50,000, the maximum mortgages then can receive is 3.5 times this amount, €175,000) and also Loan-to-value (LTV) limits. As of January 2017, first time buyers have a limit of 90% LTV on their mortgage request. Therefore, a first time buyer is required to have a 10% of their mortgage requirement as a deposit. If a person is not a first time buyer, their LTV limit is reduced to 80% consequently meaning they need 20% of their mortgage requirement as a deposit (Citizens Information, 2018). However there are some exceptions to these rules. From January 2018, financial institutions have some extra flexibility in how they can offer exceptions to potential borrowers. For LTI mortgages, they can allow up to

20% for first time buyers and 10% for second and subsequent buyer on their mortgages. For LTV mortgages, since January 2017, first time buyers are allowed up to 5% of the value of their mortgage for their primary residence in a calendar year to be above the 90% limit. For second time and subsequent buyers, the limit is 20% of the value of the new lending (Lane, 2017). All of these measures were introduced in an attempt to avoid the situation mentioned previously where the size of the mortgage market grew to over 100% of the Irish GNP.

2.2.3 Mortgage Arrears / Defaulting

As per section 3(a) of the Code of Conduct on Mortgage Arrears, a mortgage arrear is defined as:

"A mortgage arrears problem arises as soon as the borrower fails to make a mortgage repayment by the due date" (Central Bank of Ireland, 2018).

For the purpose of this research, Lender A defines a "default" as a mortgage which has been in arrears for more than 90 days.

2.2.3.1 Path into arrears

There are various reasons why a borrower may enter arrears. Generally these reasons are unforeseen and/or unavoidable and can vary wildly. For example, these reasons can range from a personal issues such as a child birth, a serious illness, a family bereavement through to issues out of the borrowers control such as losing their employment or the financial institutions raising their interest rates (Kelly M. , 2009; Norris & Brooke, 2011). (McCarthy, 2014) found that if a household has children, they are 5% more likely to enter arrears over households which have no dependent children.

Unemployment is perhaps one of the leading causes of a borrower entering arrears. (Kelly, 2009) concludes that unemployment has a far greater effect on arrears in Ireland when compared to the rest of Europe. He reasons that one crucial factor of a borrower coming out of arrears is the ability to fully service their debts and consequently they need to be in employment. Therefore he believes the focus on debt reduction is not correct and polices which look to enable the domestic economy to function correctly, and thus increase employment is far more likely to reduce the number of borrowers in arrears.

As mentioned earlier, throughout the Celtic Tiger house prices were soaring. Consequently many mortgages were taken out based on the income of two salaries to

enable the borrowers to meet large repayments. As these mortgages were dependant on two salaries, they had at a greater risk of entering arrears as it only required one of the mortgage holders to either become unemployed or have their salary reduced. Also along with this risk, another potential issue introduced with this type of mortgage is around the potential of the borrowers separating and therefore there would no-longer be two salaries to support the mortgage (McCarthy, 2014).

(Norris & Brooke, 2011) interviewed various borrowers about their mortgages and found that the majority who took out mortgages from main financial institutions felt the repayments were sustainable in the economic climate when they drew-down their mortgage, while a minority of the interviewees admitted they felt their repayment were never sustainable. From this minority of borrowers who admitted their mortgage was not sustainable, a vast majority had received their mortgage from a sub-prime lenders. Essentially, sub-prime lending is offering loans to borrowers who wouldn't be suitable for a loan from a main financial institution, for a variety of reasons such as a previous loan default or a low credit rating (Norris & Brooke, 2011). Of the borrowers interviewed some of them acknowledged they had entered arrears due to a reduction in the take home remuneration due to for example, extra levies imposed on public sector workers, however the majority of people who entered arrears due to income reduction were self-employed. Another factor identified by this research was a borrower who entered arrears due to a gambling addiction. Utilising a customer's web browsing habits could indicates dangers such as this as the customer may be moving money between their online accounts regularly, or they may be log in multiple times at hours which wouldn't be considered normal banking hours.

Along with the multitude of potential issues mentioned above, another potentially serious issue which can potentially lead a borrower into arrears are further financial commitments such as a personal loan or a credit card bill. As can be seen in table 2.1, the level of non-mortgage debt owned by Irish residents from 2000 to 2010 was a significant amount. (Keeney & O'Donnell, 2009) found that in 2007/2008, 13% of households re-mortgaged their home to unlock equity and in some cases attempt to pay-off their non-mortgage debt to enable them to only have one debt, which would be the mortgage.

Category	2000 €M	2002 €M	2004 €M	2006 €M	2008 €M	2010 €M	% Change 2000-2010
Instalment credit/hire-purchase /leases	3,337	4,720	3,595	3,659	3,651	2,461	-26%
Loans up to and including one year	23,651	21,933	19,834	32,381	58,679	55,789	136%
Other	3,901	3,027	3,723	3,861	10,440	40,748	945%
Outstanding debt on credit cards	1,079	1,461	1,898	2,407	2,820	3,034	181%
Overdrafts	7,798	7,632	6,966	7,389	8,104	8,770	12%
Repurchase agreements	172	752	921	457	1,615	763	344%
Term/ revolving loans	38,009	60,440	67,510	95,382	137,345	130,434	243%

Table 2.1 - Level of non-mortgage debt held by Irish Residents 2000-2010 (Norris & Brooke, 2011)

2.2.3.2 Dealing with arrears

The ability of a financial institution to recoup their loan is a formidable tool which can be used to dissuade a borrower from defaulting on their mortgage. This ability is called recourse and mortgages (and loans in general) are split into two categories, full recourse mortgages and non-recourse mortgages (Flavin & Connor, 2013). A full recourse mortgage gives the financial institution the ability to recover all borrowings from the sale of the security on the mortgage and also borrower themselves if the sale of the security of the loan does not cover all borrowings. A non-recourse mortgage only allows the financial institution to recover the amount from the sale of the security of the mortgage and they cannot pursue the borrowers for any potential disparity in the amount owned against the amount raised by the sale of the security (Flavin & Connor, 2013). For mortgages, the security is generally the property for which the mortgage was taken out.

In the US mortgage market, non-recourse mortgages are relatively common. In some US states, lenders may have no recourse other that their ability to reprocess and sell the property. For these states, there is an approx. 20% higher chance of default compared to states which allow full recourse (Kudlyak & Ghent, 2011). In contrast, in Ireland and the majority of Europe, financial institutions have full recourse on their mortgages and consequently borrowers are far more reluctant to default on their mortgage as they are still liable for any outstanding amount not recouped by the sale of the security of the mortgage. Due to this, the number of repositions in Ireland is relatively low when compared to the US mortgage market where repossessions are relatively common

(Kudlyak & Ghent, 2011). Financial institutions also have to abide by the new polices mentioned earlier where they have to allow a 12 month moratorium on repossessions after the first date of arrears. Finally, in Ireland repossessions have both higher legal costs and can damage a financial institutions public image. This is a further reason why Irish financial institutions are reluctant to reprocess.

The favoured technique of the Irish government, CBI and the Irish financial institutions is to attempt to restructure mortgages who have been in difficultly for a long periods (Norris & Brooke, 2011). Restructuring a mortgage can take many forms including but not limited to a switch to an interest only mortgage for a certain period of time; a reduction in monthly payment amount; a temporary deferral of payment of interest; arrears capitalisation (spreading the arrears amount over the rest of the term of the mortgage); extending the term of the mortgage; a write-down of some of the outstanding balance or a split mortgage (the mortgage is split into two parts to enable the reduction of the monthly repayments (Banking & Payments Federation Ireland, 2015). As of December 2017, 118,477 PDH mortgages were considered as restructured. Figure 2.5 shows the distribution of options of restructuring and clearly shows that arrears capitalisation and split mortgages are the most common method of restructuring mortgages accounting for approx. 56% of all mortgage restructures at the end of December 2017.

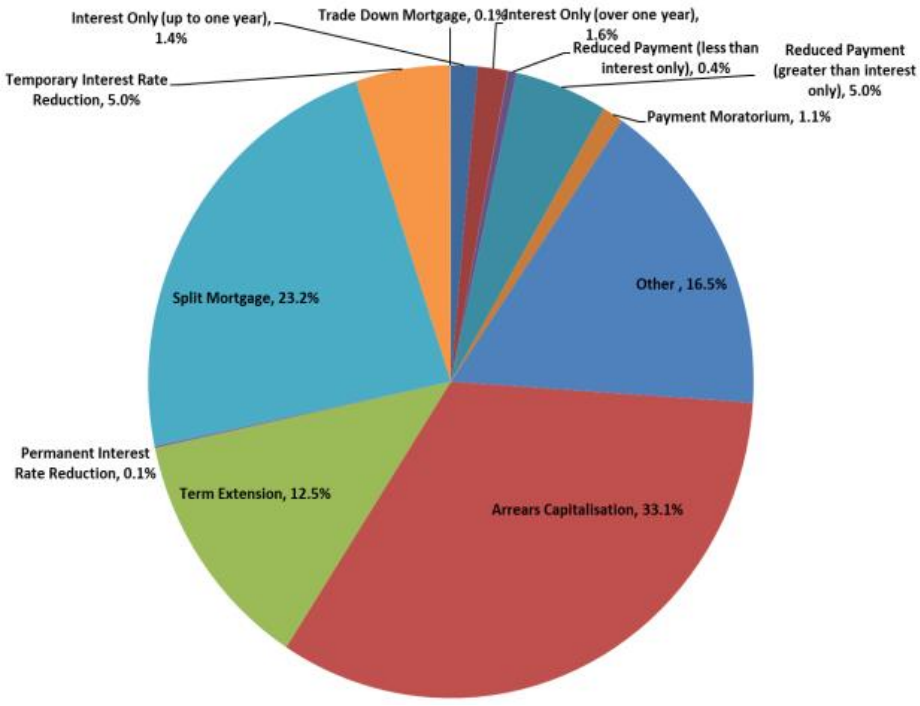

Figure 2.5 - Percent of restructure option Dec 2017 (Central Bank of Ireland, 2018)

As of December 2017, there was 729,722 PDH mortgages in Ireland, equating to approx. €98.5 billion. Of these 70,488, or approx. 10% were in arrears. Of these mortgages in arrears, 48,433 were in arrears for more than 90 days. For these account in arrears over 90 days, they equate to approx. 10% of all outstanding balances for PDH mortgages (€9.7 billion). If these accounts were held by Lender A, they would therefore be considered as having defaulted. The number of mortgages in arrears has been declining steadily since mid-2013, which is a positive trend and can be seen in figure 2.6.

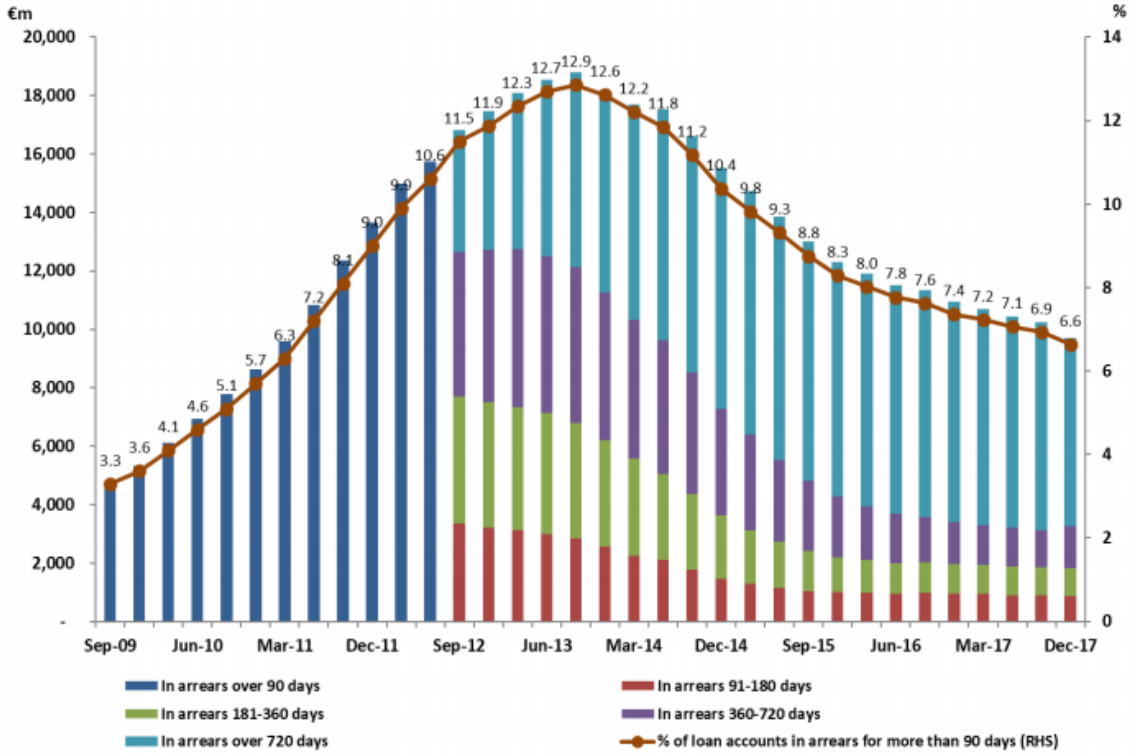

Figure 2.6 - Number of mortgages in arrears 2009 – 2017 (Central Bank of Ireland, 2018)

2.3 Clickstream Data

Clickstream data are an extremely powerful data source which is used extensively in marketing and Customer Relationship Management (CRM). (Poel & Buckinx, 2005) utilises clickstream data to predict purchasing behaviour of a user, while (Raphaeli, Goldstein, & Fink, 2017) utilises clickstream data across various channel's such as web and mobile to determine how users are interacting with websites and conclude that users of the website are more exploratory while mobile use is more task focused. This one key insight alone could help drive an organisation's decision on the information to display on their website for users who are accessing with different methods and devices.

(Bucklin & Sismeiro, 2009) defined clickstream data as "*the electronic record of Internet usage collected by Web servers or third-party services*" (Bucklin & Sismeiro, 2009). Another potential way of thinking about clickstream data is to consider it an activity log of a user's interactions with a digital service, be it a website, mobile/tablet app or a bank branch kiosk for example. Clickstream data can be collected in a variety of methods, but all methods fall into one of two categories: site-centric or user centric. Site-centric occurs when an organisation monitors the users of one specific website and user-centric occurs when a user is monitored across various websites.

Site-centric monitoring is generally used by smaller organisations who's users may be paying for a service, in the case of this research Lender A's customers who interact with the digital services they offer. User-centric would be used by organisations who drive profit by displaying relevant ads to it users. To enable them to display relevant ads, they track all aspects of customer while they browse the internet. For example, Google will store all searches a user makes and will also store any links a user's clicks on from the search (Tene, 2008). This enables them to get detailed information about a user without the user even realising it. If a person was to search for "how to trim dog nails", Google could infer the person has a dog and consequently display ads for pet groomers local to the person as their general location can be determined via the IP address.

In addition to site-centric and user-centric collection methods, the actual data collection falls into two categories: server-side collection and user-side collection (Srivastava, Cooley, Deshpande, & Tan, 2000). (Srivastava, Cooley, Deshpande, & Tan, 2000) conclude that user-side collection is a superior method as it alleviates some of the issues related to server-side collection, such as when pages are cached and do not require a call to the server. However one pitfall of user client-side collection is the ability of the client to disable the collection of data if they wish. An example of this would be if an organisation is employing JavaScript to perform the monitoring and collection, a user could disable JavaScript or simply block the script which is executing the monitoring and consequently the organisation would have no collected data for that user. If this occurs, the organisation can implement changes to force the user to enable JavaScript, but this may drive the user away from the digital service.

Lender A employs various technologies to enable them to monitor their customers as they interact with their digital channels which enables them to get a deeper

understanding of their customers by understanding how they interact with their digital channels. They capture information such as the device being used, time of logon, what kind of actions the user is making (such as transferring money etc.), how long they spent logged on, what specific pages they visited while on the website amongst other information. Currently, this information is being used to drive email campaigns to customers by enabling Lender A to understand who is interacting with their emails. There is also on-going work to determine how valuable clickstream data is when used for fraud detection with early indications showing it can provide a positive outcome for fraud detection.

2.4 Data Mining

As technological advancements continue at a blistering pace and it's ubiquity in our everyday lives also swells, more and more organisations are realising the potential beneficial value of effectively leveraging their customer's data for various reasons such as profit growth and improving the CRM with their customers as well as protecting themselves. As an example of this, Amazon.com offers personalised home pages to its customers based on personalisation models and many credit card and phone companies use the analysis of their customer's data to enable fraud identification (Yang & Padmanabhan, 2003). The use of these models would fall under the umbrella term of Data Mining which is the process of examining and analysing existing data in an attempt to discover new useful knowledge.

(Frawley, Piatetsky-Shapiro, & Matheus, 1992) proposed a process of enabling the discovery of finding useful knowledge in data, known as 'Knowledge Discovering in Databases' (KDD). Data Mining is one of the core components of KDD, but it is only one of a number of steps including selection, pre-processing and transformation amongst others. They defined KDD as

> *"The non-trivial process of identifying valid, novel, potentially useful, and ultimately understandable patterns in data."* (Frawley, Piatetsky-Shapiro, & Matheus, 1992)

Within the discipline of Data Mining, there are several different statistical and analytical approaches which can be used to extract knowledge from data. Perhaps one of the more educational visualisations to represent Data Mining was created by Dr Syed Sayad, a professor from the University of Toronto with over 20 years knowledge in data science, machine learning and artificial intelligence (Sayad, 2018). Dr Sayad created the "Data

Mining mind map which shows various techniques using during the Data Mining process and can be seen in figure 2.7.

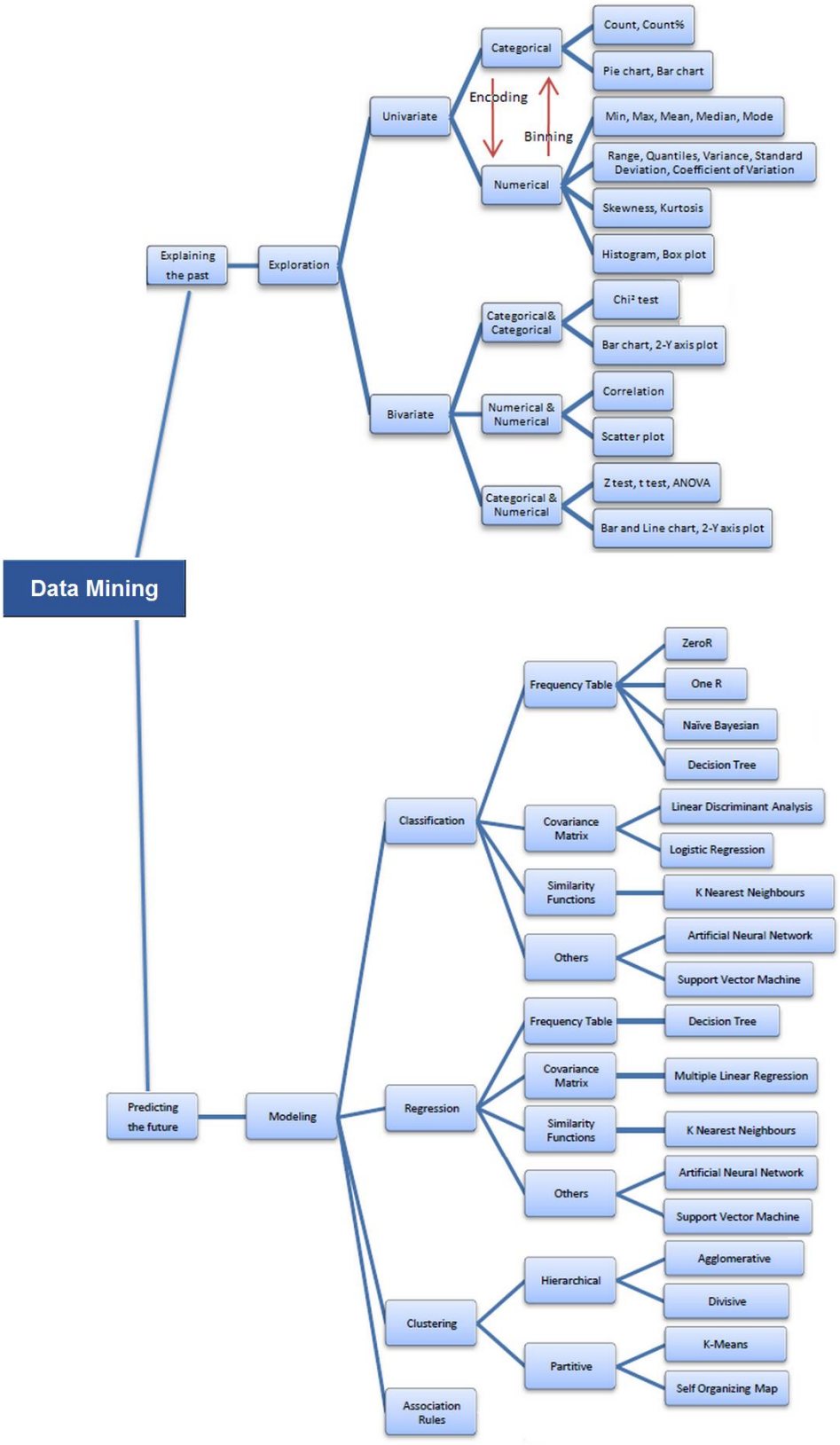

Figure 2.7 - Data Mining Map

2.4.1 Predictive Modelling

For this research, the main interest in a subset of Data Mining known as *Predictive Modelling*. Predictive modelling is a method where-by there is an attempt to predict future events based on historical events. To achieve this, historical samples containing explanatory attributes relative to a target outcome which we are attempting to predict are investigated to determine their correlation. The models used for this process function by accepting these attributes as input and applying mathematical transformations in an attempt to predict what the outcome associated with the input will be. As historical events are employed to build the models, it enables the accuracy of the model to be easily validated as the events used to construct the models can be examined at a later date to the events being used to train and test the model.

One of the key challenging factors associated with attempting to use predictive modelling to predict future events is the attributes of these unseen events are not likely to be the same as the events which are being utilized to train the model. For this reason, it is crucially important that the model is not over-fit to the historical events as this may not lead to a model that will generalise well to future events. Certain processes and methods can be employed in an attempt to reduce the possibility of this occurring and these will be discussed further in subsection *2.7*.

2.5 Modelling Algorithms

The goal of this research is to determine if a customer's clickstream data can be utilized in addition to their demographic and transactional data to determine if they are at risk of entering arrears on their mortgage. As mentioned above, predictive modelling is the attempt to predict future events based on historic events. To achieve this, one or more modelling algorithms are used. Predictive modelling can be applied in various locations across broad spectrums of technological and business area such as cancer prediction (Yokota, et al., 2003), student attrition (Camba, David, Betan, Lagman, & Caro, 2016), bankruptcy prediction (Chaudhuri & De, 2011) and reading car number plates (Li, Wang, You, & Shen, 2018).

When applied correctly, predictive modelling can be extremely beneficial to an organisation. However, if an organisation doesn't employ a suitable algorithm for their data, they may not see favourable results from their modelling. Financial institutions have been using models for many years as an attempt to reduce their overall risk and

drive their profits upwards. Since the economic meltdown of 2008, this has become even more important to the financial sector as they attempt to avoid a repeat of the 2008 crash.

With the tightening of the Irish and EU credit law, both Irish and EU regulators are now very stringent to ensure EU financial institutions are complying with the latest IFRS9 (International Financial Reporting Standard 9) accountancy standardisation rules and are continually auditing EU financial institutions to enforce the IFRS9 rules. IFRS9 sets out guidelines which financial institutions must follow and it specifics what models a financial institution must have, such as a model for Expected Loss Impairment (Leman, 2015).

There are many algorithms which could be employed for both the requirements of IFRS9 and other predictive modelling requirements of Lender A, such as mortgage arrears prediction. In this section, five algorithms will be examined closely to determine how they function and their potential uses.

2.5.1 Decisions Trees

Decision Tree algorithms function by attempting to split the data into classes based on the attributes contained within the data and that attributes relationship to the target attribute (class). Preceding the splitting of the data, if a record was randomly chosen from the dataset, the degree of ambiguity as to which class it will belong to is entirely relative to the distribution of classes in the original dataset. For example, if a random record was chosen from the table 2.2, there would a 30% chance the piece of fruit was edible, and a 70% chance it is not edible. This is because there were 7 of the 10 fruits which are not edible while only 3 are.

Colour	Size	Edible
Green	Small	No
Red	Small	Yes
Green	Small	No
Green	Small	No
Green	Large	Yes
Red	Small	No
Green	Large	No
Green	Large	Yes
Red	Small	No
Red	Small	No

Table 2.2 - Sample Dataset

The goal of the decision tree is to reduce the ambiguity of which class a customer would belong to if one was chosen at random. (Shannon, 1948) states that in the above scenario, if five fruits were edible and five were not, then there is the maximum degree of uncertainty in the dataset. This uncertainty is denoted by the symbol *H* and is called Entropy (Shannon, 1948), which is measured in *bits.* The formula for entropy is as follows:

$$H = -\sum_{i} p_i \, log_i \, p_i$$

After the splitting of the dataset, the decrease in the degree of uncertainty is known as the Information Gain, denoted as *IG* (Quinlan, 1986). The formula for Information Gain is as follows:

$$IG = Entropy\ before\ split - Entropy\ after\ split$$

To allow for the creation of a new *leaf* node, the dataset is split using the available features. Looking back at table 2.2, if the attribute "Size" is used to split the data it will create two leaf nodes, one for small fruit and one for large. In the "small" node, there will be only be one edible fruit out of six, while in the "large" node, two out of the three fruits will be edible. Therefore if a fruit was chosen at random from the "large" node, there would be a 66% chance it would be an edible fruit, while there would only be an approx. 15% chance of randomly picking an edible fruit from the "small" node. For the "large" node, the degree of uncertainty of choosing an edible fruit has gone from 30% in the root node to 33%, an increase of 3%. Conversely, if you look at the "small" node, you have an approx. 15% chance picking an edible fruit at random.

Using table 2.2, the entropy of the "Edible" attribute is 0.881 bits with the entropy of "Size=small" being 0.592 bits and "Size=large" being 0.918 bits, and therefore the entropy of the "Size" attribute is 0.690 bits. Consequently, the entropy has gone from 0.881 bits to 0.690 bits. Therefore, the Information Gain for the "Size" attribute is 0.191 bits (0.881-0.690). This process is then repeated for the remaining attribute, "Colour". If there is an Information Gain, new leaf nodes can be created. The outcome of this decision tree can be seen in figure 2.8.

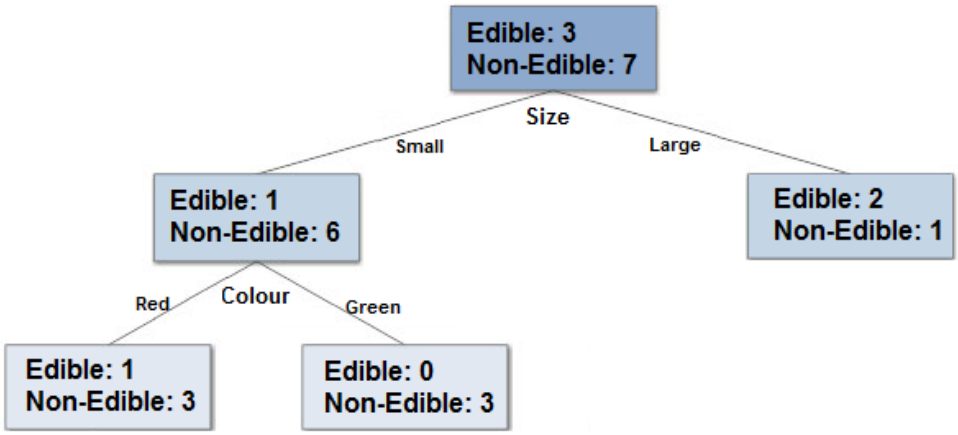

Figure 2.8 - Visualisation of decision tree created for example

A well-documented potential issue with decision tree is where by two records in the dataset have the same attributes but different classes (Quinlan, 1986). If this issue occurs the decision tree will not be able to confidently determine which class the record should belong to. Consequently the decision tree will assign the class of the majority class within that group (Kohavi , Yun , & Friedman , 1996). If the situation persists and there is still a tie between the classes, then the class is decided based on the majority class of the parent note from which the subgroup was formed.

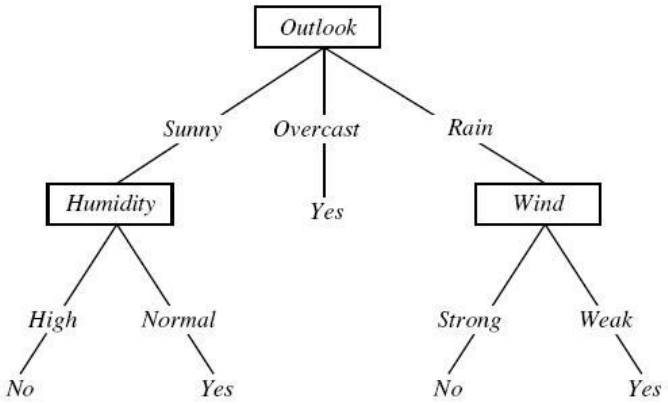

Figure 2.9 - Basic decision tree (Quinlan, 1986)

Due to how decision trees are built, once there is sufficient information in the underlying data to make a split, the tree will continue to grow. As can be seen in figure 2.9 and 2.10, both of these trees correctly classify a training dataset into 'P' and 'N'. In this scenario, (Quinlan, 1986) suggests that the simpler of the trees, figure 2.9, should be chosen. There are two main reasons for this, firstly performance and secondly over-fitting (Ho, 1995). The second tree in figure 2.10 is much larger and will therefore require far more

computational power to build compared to the smaller in figure 2.9. Due to the complexity of the second tree, it is potentially liable to over-fitting to the training data. Consequently while both trees correctly classify the training data, the more complex tree may struggle to correct classify unseen records.

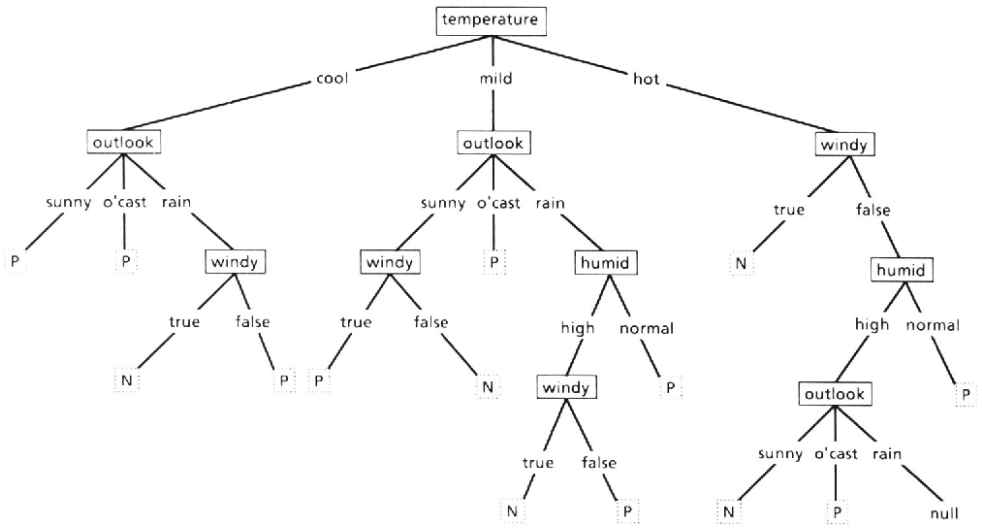

Figure 2.10 - Complex decision tree (Quinlan, 1986)

As mentioned earlier, Information Gain can be a method used to determine the best splitting location during the construction of a tree. However a distinct disadvantage of Information Gain in the potential for it to lead to over-fitting. As Information Gain has a preference of creating small splits in the data, it can have detrimental impacts when an attribute such as a unique id is introduced. With this field, the algorithm would correctly split the data on each of the unique ids' and create tree with a size of 1 and an entropy of 0. However this tree could not be deployed as it would not have the ability to classify previously unseen unique id's presented to it. To mitigate this issue, (Quinlan, 1986) suggests to use the measure of Gain Ratio which is the Information Gain divided by the Entropy after the Split. (Quinlan, 1986) found the utilising Gain Ratio as the splitting method often produced more favourable results over Information Gain. For the above example of unique id's, if Gain Ratio was used the *entropy after the split* would be 0 and consequently the Gain Ratio would be 0. Therefore other attributes in the dataset would return better Gain Ratio's and would be chosen as the splitting attribute over the unique id attribute.

As well as Gain Ratio, another measure which could be used to alleviate Information Gain predilection for attributes with many values is *Symmetric uncertainty*. This

measure functions by normalising the data to a range of 0 to 1 and is not influenced by attributes with multiple values (Liu & Yu, 2003). In addition to all of the measures already mentioned, another measure which could be used is Gini Index. Gini Index measures the impurity of a node and is a measure of how often a randomly chosen record from a dataset would be incorrectly classified if it was randomly classified according to the distribution of classes in the dataset. In addition to being used to build decision trees, the above measures of Information Gain, Gain Ratio, Symmetric uncertainty and Gini-Index can also be used during the feature selection process.

For simpler predictive models, decision tress are generally good models as they are easy to explain and interpret. However as the complexity grows, the effectiveness of a singular decision tree begins to weaken in both accuracy and effectiveness (Kohavi , Yun , & Friedman , 1996). As discussed by (Kohavi , Yun , & Friedman , 1996), decision trees can be susceptible to various issues and therefore this leads to them rarely being used for mortgage arrears prediction. One of these crucial issues is due to how decisions trees are built. As they are constructed with a top down approach, they can only ever look one step ahead. Consequently the algorithm may not chose the optimal split for the node as each attributed is treated independently to the rest of the attributes. For larger datasets with multiple attributes, it would not be uncommon for one attribute to have little correlation to the target attribute, while a combination of attributes may be a strong predictor. Therefore, using multiple decision tress should perform better than a singular tree. This is where the concept of random forests come from which are discussed in detail in the following section.

2.5.2 Random forests

First proposed by Tin Kim Ho at AT&T Bell Labs (Ho, 1995) and later refined by Leo Breiman and Adele Cutler in 2001 (Breiman & Cutler, 2001), Random Forests (RF) are an ensemble learning method for classification. RF are popular and accurate algorithms and are therefore commonly used as the default classification models in many applications (Zhang, Liu, Zhang, & Almpanidis, 2017). They were originally suggested and developed to alleviate the potential issue of single decision tree's habit of over-fitting their training data and not generalising well to unseen data.

RF operate by constructing a mass of random un-pruned tress and using the mode of the output of all of the tress as the output of the algorithm (Ali, Khan, Ahmad, & Maqsood,

2012). Trees in Random Forests are constructed using the following method (Breiman & Cutler, 2001):

1. We first need to determine which sample data will be used to build the tress in the forest. For this, if we say that N is the number of cases in the training dataset, we sample N cases at random, but with replacement from the original dataset. This will be the sample which will be used as the training dataset for growing the tree.
2. If there are M input attributes, a number of m<<M is defined such that at each node, m attributes are selected at random from M and the best split from these is used to split that node. During the growth of the forest, the value of m is held constant.
3. Each tree is grown to the largest possible extent. No trees are pruned during the process.

Two effects which can impact the error rate of RF are correlation between two trees and the strength of the individual tress of the forest. These effects can be alleviated to improve the error rate by adjusting the *m* value during the construction of the individual trees (Breiman & Cutler, 2001).

In 2012, (Ali, Khan, Ahmad, & Maqsood, 2012) performed a comprehensive analysis comparing Random Forests to singular decision tress on breast cancer data and concluded that "*Random Forest achieves increased classification performance and yields results that are accurate and precise in the cases of large number of instances*" (Ali, Khan, Ahmad, & Maqsood, 2012). They also discussed random forests ability to handle missing values and their unlikeliness to develop over-fitting problems due to the missing values. In their comparison of state-of-the-art classification algorithms, (Zhang, Liu, Zhang, & Almpanidis, 2017) concluded that random forests proved to be the joint highest performing classifier (along with Gradient Boosting Decision Trees), closely followed by support vector machines, which will be discussed later in subsection *2.5.5*.

In addition to this, (Pal, 2005) validated the effectiveness of random forests when he compared them to support vector machines and concluded the performance of both was comparable. As the problem we are attempting to determine with this paper is whether someone will enter arrears on their mortgage which is a binary classification, random forests would be a suitable classifier to employ. (Barboza, Kimura, & Altman, 2017)

further confirmed the power of RF when the concluded RF was the highest preforming model for bankruptcy prediction, achieving better results than neural networks and support vector machines, which can be seen in table 2.3.

Model	TP	TN	FP	FN	Type I Error (%)	Type II Error (%)	AUC (%)	ACC (%)
SVM - Linear	141	9,203	5,350	6	4.08	36.76	66.31	63.56
SVM - RBF	129	10,900	3,653	18	12.24	25.01	89.54	75.03
Boosting	113	12,575	1,978	34	23.13	13.59	91.25	86.31
Bagging	116	12,452	2,101	31	21.09	14.44	91.49	85.5
Random forest	112	12,536	2,017	35	23.81	13.86	91.15	86.04
Neural networks	138	10,047	4,506	9	6.12	30.96	91.09	69.29
Logit	133	9,659	4,894	14	9.52	33.63	86.18	66.61
MDA	119	7,254	7,299	28	19.05	50.15	67.4	50.16

Table 2.3 - Model performance from (Barboza, Kimura, & Altman, 2017)

2.5.3 Regression

Regression is used to model linear relationships between attributes in a dataset. In its most basic form, a linear regression model can be thought of as an attempt to predict the change of one dependant variable based on the change of another independent variable. The potential models which can be built in this scenario can be represented as $Y=F(x)$ where Y is the dependant variable we are attempting to predict by applying some function to the independent variable x. A simple example of this model can be seen in figure 2.11 where a person's salary increases as their number of years of experience increases. This model can be represented as $Y = x$, where as a person's years of experience increases, their salary also increases.

Figure 2.11 - Simple regression line example

Obviously, this is the simplest form of regression and the majority of real world scenarios are not as easy to understand and visualise. The majority of real world

scenarios will require fitting relationships there are polynomial or curved. This type of relationship can be seen in figure 2.12, where (Mehdikarimi, Norris, & Stalzer, 2015) are modelling the income levels compared to the number of hours worked in the US based on 2013 Census data which showed that as wages per hour increases past a certain point, the number of hours worked decreases when compared to those on a lower pay scale.

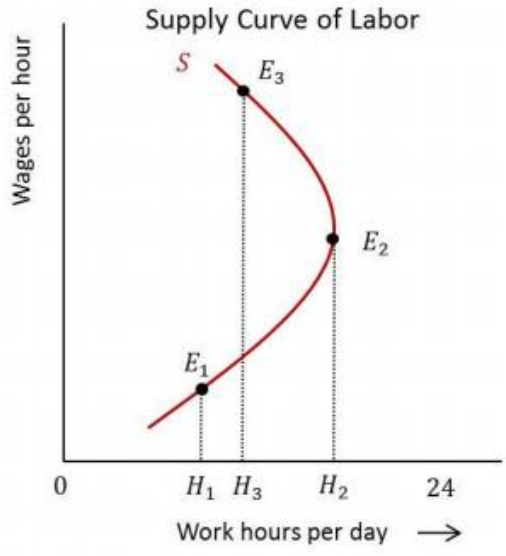

Figure 2.12 - Complex regression line example

When attempting to predict a binary outcome, such as whether a customer will enter arrears or not, a form of regression called logistic regression is more commonly used (Cai & Guo, 2016). Due to its relative ease of interpretation and the fact it is an effective method for the classification of binary outcomes, logistic regression is widely used across various different sectors (Mehdikarimi, Norris, & Stalzer, 2015; Hariri-Ardebili & Pourkamali-Anaraki, 2018; Angelini, 2018).

2.5.4 Neural Networks

The concept of a neural network was first conceived in the 1940s in a research article by McCulloch and Pitts in which they showed in principle that even simple types of neural networks could compute any logical or arithmetic function (McCulloch & Pitts, 1943). The human brain is an extremely complex and powerful machine which is excellent at solving complex problems at blistering speeds when compared to computers. Conventional computers perform well while processing large amounts of data but are very limited in the processing they can perform as they can only preform functions for which they were coded or from which they learned in their past activities.

The human brain solves problems by summing both electrical and chemical signals transmitted by *synapses* into *neurons* (Huang & Luo, 2015). From here, the neuron takes the inputs from the synapses and they decide whether or not to output an output signal to a further neuron based on whether the summation of the input signals passes a specific threshold. Thus, the neuron exhibits an all or nothing property (Huang & Luo, 2015). The brain is made up of billions of neurons interconnected to each other, and this is basically how the human brain functions (Huang & Luo, 2015).

Neural networks attempt to mimic the functions of the human brain when attempting to solve problems (Lapedes & Farber, 1988). The architecture of a traditional neural network consists of have one input layer, one or more hidden layers and an output layer, which can be seen in figure 2.13 which is a traditional neural network consisting of two hidden layers. The output layer is attached to every hidden layer node and some of the input nodes. The hidden layer nodes are connected to one or many other hidden layer nodes, the output node and one or multiple input nodes. Each of the connections between the nodes are weighted and therefore this allows for the output of the model to be modified to a certain extent and also allows for the network to learn from the data.

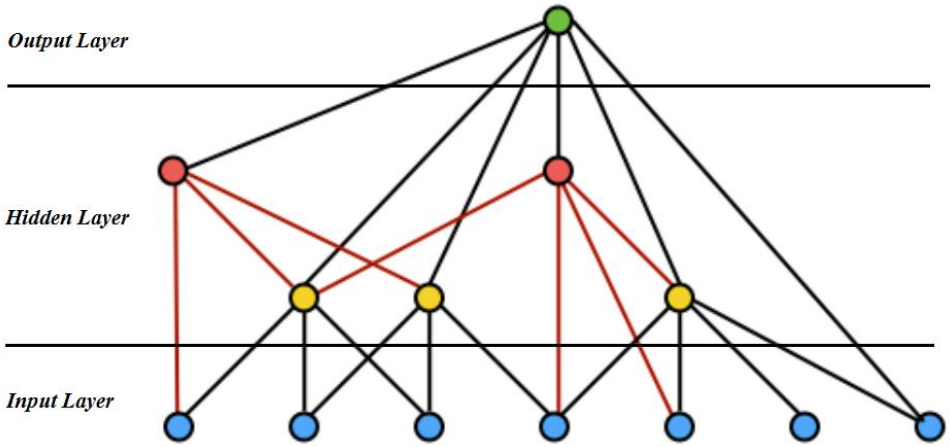

Figure 2.13 - Typical neural network architecture (Cortes, Gonzalvo, Kuznetsov, Mohri, & Yang, 2017)

The hidden layer is the portion of the neural network which tries to mimic the human brain. As the name implies, it is extremely difficult and in many cases impossible to understand how the hidden layer is performing its calculations and therefore it may be impossible to determine why a specific set of input attributes results in a specific output. Consequently, while neural network have been shown to perform well at various tasks such as predicting credit card delinquents (Hao & Zhang, 2009), they are not the first

choice for financial institutions for default prediction as the financial institutions may need to explain how they work to a financial regulator.

For example if a financial institution decided, through a neural network, a customer was going to default on their loan and then contacted the customer, and it's turns out the customer is not at danger of default the customer may become offended and contact the Central Bank of Ireland or the Financial Ombudsman and lodge a complaint against the financial institution regarding the issue. If they then decide to investigate why the customer was classified as a default risk, the financial institution may find it extremely difficult to explain their neural network. However, (Hara & Hayashi, 2015) demonstrates this problem can somewhat be alleviated by utilising a rule extraction algorithm called Re-RX. This allows for the rules of the neural network to be extracted similar to a decision tree and therefore conclusions can be drawn to understand how a neural network preformed its function and decided on a specific outcome. Nevertheless, unlike a decision tree, these rules can be very complex and problematic to understand (Hara & Hayashi, 2015).

2.5.5 Support Vector Machines

Originally developed for binary classification in 1995 by (Vapnik & Cortes, 1995), Support Vector Machines (SVM) are generally recognised to be one of the most reliable and accurate algorithms in predictive modelling (Hariri-Ardebili & Pourkamali-Anaraki, 2018; Zhang, Liu, Zhang, & Almpanidis, 2017). SVM algorithms attempt to split the sample data into distinct classes via a linear Hyperplane. It achieves this by attempting to maximise the *margin* between the nearest *support vectors*. This can be seen in figure 2.14 where there are two distinct classes linearly separated.

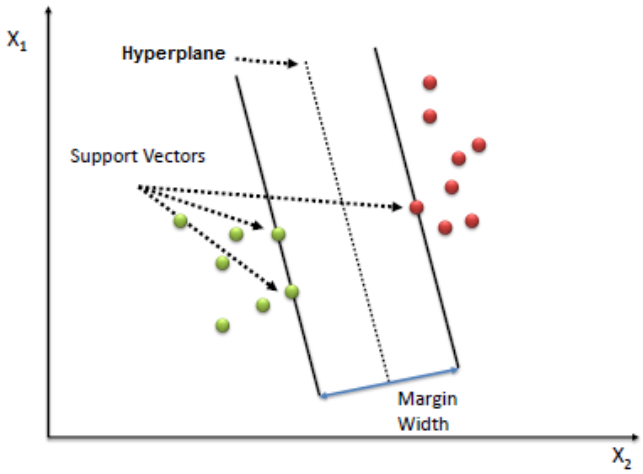

Figure 2.14 - Sample SVM visualisation

The data in figure 2.14 is ideally split to allow for SVM to function perfectly as it is linear split, however what happens in the scenario where the data is not linearly separated such as in figure 2.14? As can be seen in figure 2.15, there may be no location to draw the Hyperplane and consequently it may seem as though SVM cannot be applied to this data.

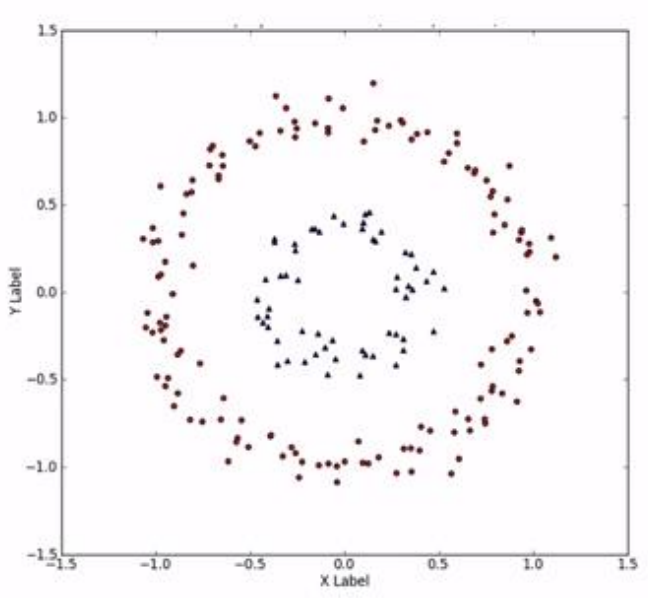

Figure 2.15 - Complex dataset for SVM in one dimension

However if another dimension was added to the data, it may allow for a hyperplane to be drawn, as can be seen in figure 2.16. To achieve this modelling on non-linear data, SVM uses what is commonly called a *Kernel Method* (sometimes also referred to as *Kernel Trick*). In terms of SVM and Machine Learning, a kernel is essentially a mapping function with the ability transform a given space into a different space, usually a high dimensionality space.

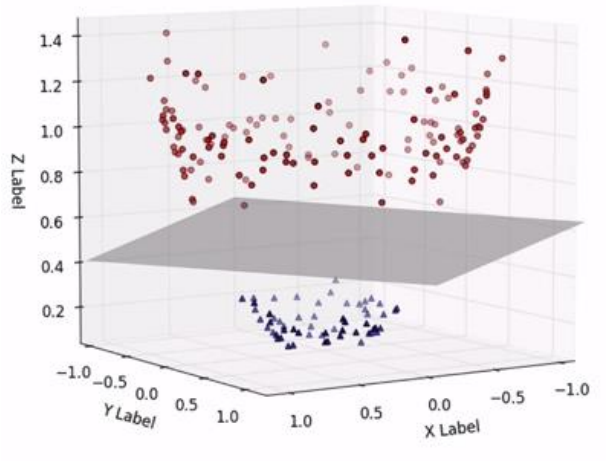

Figure 2.16 - Complex dataset with extra dimension added

Simply put, a kernel method adds more dimensionality to the data in an attempt to allow the SVM to find a suitable hyperplane which splits the data accurately (Zoppis & Riccardo Dondi, 2018). The kernel method can use a variety of options in an attempt to allow the SVM to find a hyperplane. These options include linear, radial basis function (RBF) and polynomial amongst others.

As with all predictive modelling algorithms, SVM are susceptible to misclassification errors. An example of this can be seen in figure 2.17, where one red example has been misclassified as a blue example. In SVM, misclassification rates are controlled by the *C* parameter (Cherkassky & Ma, 2004).

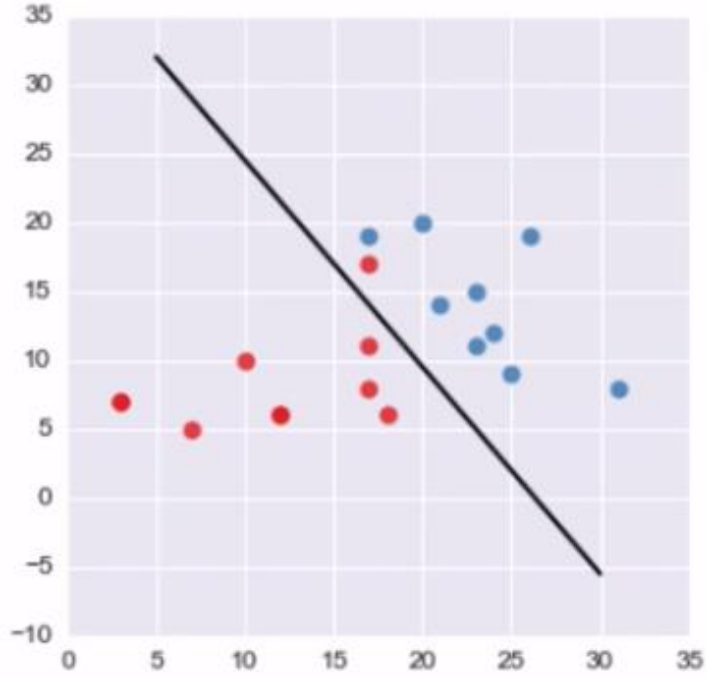

Figure 2.17 - Simple SVM output showing misclassified sample

The *C* parameters controls the trade-off between a smooth decision boundary (margin) and classifying the training points correctly. In figure 2.17, it is assumed the *C* value is 2^{-5}. If the *C value* was amended so that it was increased to 2^{5}, the resultant scatterplot can be seen below in figure 2.18. While this model is now correctly classifying all of the data, it strongly seems to be over-fitting the data and may not generalise well to unseen data (Rakotomamonjy, 2004). When choosing a value for *C*, careful consideration must be taken to reduce the possibility of over-fitting the training dataset.

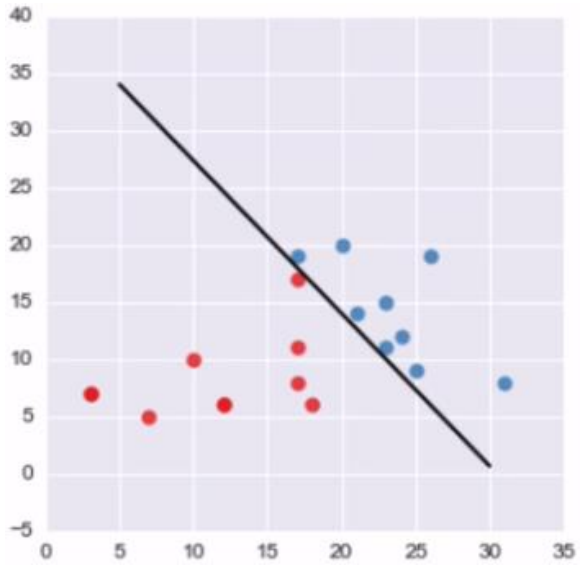

Figure 2.18 - Simple SVM output showing potential over-fitted model

SVM have proven themselves of be an accurate classifier and often outperform other classification algorithms in various areas such as optical character recognition (Phangtriastu, Harefa, & Felita Tanoto, 2017), student attrition (Camba, David, Betan, Lagman, & Caro, 2016), predicting diabetes (Zhang, Liu, Zhang, & Almpanidis, 2017) and determining the reliability of concrete dams (Hariri-Ardebili & Pourkamali-Anaraki, 2018) to name just a few. They have also been showed to be useful for the purpose of bankruptcy prediction with (Chaudhuri & De, 2011) finding them to perform better than neural networks and logistic regression as can be seen in table 2.4. Like random forests perform better than singular trees, ensemble SVMs have been demonstrated to generally perform better than single SVMs, however, generating multiples SVMs will be extremely computational expensive.

Average prediction accuracy; training and validation figures are expressed in percentage.

	LR		ANN		Pure SVM		GA-SVM		FSVM	
	Training	Validation	Training	Validation	Training	Validation	Training	Validation	Training	Validation
32FS	78.13	68.18	79.57	68.18	82.45	72.22	86.53	80.30	87.69	83.37
30FS	80.53	67.68	78.85	69.19	84.86	71.72				
12FS	66.83	68.69	79.81	69.19	81.73	72.73				
6FS	76.92	70.71	75.48	71.72	81.01	74.75				

Table 2.4 - Comparison of various models (Chaudhuri & De, Fuzzy Support Vector Machine for bankruptcy prediction, 2011)

Based on this, it seems likely the SVM would perform well in an attempt to predict mortgage arrears and default classification, however careful consideration will have to observed as attempting to model data with a multitude of features can be extremely computational expensive as the SVM attempts to build a *N* dimensional space to allow it to determine the best location for the hyperplane (Joachims, Finley, & John Yu, 2009).

As mentioned earlier, in their comparison of state-of-the-art classifiers, SVM was only outperformed by gradient boosting decision trees and random forests. They concluded SVM performed well across all of their tests and also stated SVM was, along with K-NN, the most time efficient classifier during their testing (Zhang, Liu, Zhang, & Almpanidis, 2017).

2.6 Skewed Datasets

Skewed datasets, also called imbalanced datasets, can cause issues during various predictive modelling problems and thus can lead to a decrease in the accuracy and/or effectiveness of the model. A dataset is consider skewed/imbalanced when the class of interest is relatively rare compared to the others classes of the dataset (Hoens & Chawla, 2013). It would not be uncommon for the focus of the modelling to be the minority class of the dataset and consequently this can lead to issues.

For example, as of December 2017, there was 729,722 PDH mortgages held in the entire Irish mortgage market, with approx. 7% of these in arrears over 90 days (48,433) and consequently can be considered as defaulted mortgages (Central Bank of Ireland, 2018). If a model was built and just predicted all of the market stock would not default, it would be correct 93% of the time and therefore have an accuracy of 93%. Thus if a model was to be built to predict mortgage defaulting across the market, something needs to be done about this heavily imbalanced dataset.

If the model mentioned above was used, it would appear to be a high preforming model, but obviously this would not be the case. While there would be little issue with predicting a preforming mortgage will default, there would be a risk associated with classifying a mortgage at risk of default as a preforming mortgage. This 7% of the market in arrears accounts for approx. €9.7 billion of outstanding PDH mortgage balances, which is 10% of all outstanding PDH mortgage balances across the country. If these mortgages were flagged as being in a good state, future mortgages likely to default could be incorrectly classified and therefore the risk of serious repercussions to both the financial institutions holding these mortgages and also the Irish economy could be substantial.

(Hoens & Chawla, 2013) state a dataset should be split 50/50 across the majority and minority classes to avoid the issues mentioned above. However in reality this is rarely the case. The majority of prediction and classification models are designed to function correctly with balanced datasets and thus the models may become bias towards the

majority class of the dataset (Longadge, Dongre, & Malik, 2013). To counteract this issue, a number of potential methods can be employed to balance the dataset used for development and validation of the model. Some of these methods are discussed further below.

2.6.1 Random Sampling

The most common random sampling technique used is a Simple Random Sample (SRS). Simple random sampling functions by randomly picking *n* number of samples from the full dataset and using the sub-dataset as the dataset used during the modelling. SRS does not address the issue of imbalanced datasets, and may in fact lead to an even more imbalanced dataset. Therefore, other random sampling techniques should be employed to avoid imbalanced sample datasets.

2.6.1.1 Undersampling / Oversampling

Two further random sampling techniques which fall under the topic of random sampling are random undersampling and random oversampling. They both attempt to alleviate the issue of skewed datasets by either reducing the number of samples of the majority class in the dataset or by increasing the number of samples of the minority class of the dataset (Hoens & Chawla, 2013). Undersampling functions extract a random sample of the majority class of the dataset and uses this to construct the model in place of the full majority dataset. Consequently this can balance the dataset. Oversampling is basically the opposite of undersampling and functions by randomly extracting minority samples from the dataset and re-adding them to the dataset in addition to the original dataset. Therefore this will duplicate some of the minority class of the dataset.

Undersampling decreases the overall number of samples in the dataset, while oversampling will lead to an increase in the number of samples. Both of these techniques can improve the split of the minority/majority class of a dataset, but they can also introduce their own issues (Longadge, Dongre, & Malik, 2013). Using undersampling, it is possible to remove important samples from the dataset which may lead to important factors being missed by the model. Conversely, oversampling can lead to a model becoming over-fit to the minority class (Hoens & Chawla, 2013).

2.6.2 Stratified Sampling

Stratified sampling can also be used to alleviate the issue of imbalanced datasets. Stratified sampling functions by splitting the dataset in smaller datasets called stratum (plural: strata). From here, an SRS is applied and a certain number of samples are drawn

from each strata. Consequently the final sample dataset will contain an even number of samples from the majority and minority classes of the original dataset.

For a mortgage arrears dataset, this would involve creating two strata. One containing the mortgages which are preforming and not in arrears and one containing mortgages which are in arrears. From here, a SRS would be applied on *n* number of samples which will be selected from each strata and the result of both the SRS's applied to both strata will be used as the dataset to build the model.

2.6.3 Synthetic Sampling

One final method of sampling which can be used to address the imbalance of a dataset is synthetic sampling. Synthetic sample is similar to random oversampling in that it adds rows from the minority class to the dataset, but it does not duplicates the rows. It functions by creating new rows which have similar attributes to the already existing minority class rows and adds these rows to the dataset. It does this by creating a sample between a minority class sample and its nearest neighbour (Longadge, Dongre, & Malik, 2013). One common method of synthetic sampling is synthetic minority oversampling technique (SMOTE).

2.7 Model Evaluation

Evaluating the performance of a model is perhaps one of the most crucial steps in the data mining journey. When a model is being evaluated to confirm its predictive power, there are some considerations which need to be made to enable the model to perform to the best of its ability. If a model is built using a full historical dataset and then validated against the same dataset, there is a very likely possibility the model will over-fit the data and will be biased towards the data. If this occurs, the model will not generalise well to unseen data and consequently will be a poor model. This method of training and validation is sometimes referred to as *Resubstitution Validation* and it obviously not a good methodology to follow (Dougherty, Hua, & Bittner, 2007).

Reducing the bias of a model to its training dataset is an important step in the model development and evaluation. A method which can be employed to reduce this bias is the *holdout method*. This is possibly one of the most common validation methods used (Schorfheide & Wolpin, 2011). It involves splitting the data set into a training and validation datasets and training the model on the training dataset and validating it on the validation dataset. A split of 70/30 for training/validation is commonly used when the holdout method is employed (Schorfheide & Wolpin, 2011). Along with the holdout

method of validation, another common method used is the Cross-validation method. Sometimes known as K-Fold Cross-validation, this is a method where the dataset is split into K distinct equally sized samples (folds) (Rodrígues, Pérez, & Lozano, 2010). From here, the model is trained on k-1 folds and validated against the fold left out. The model is trained k times and uses the folds left out each time for validation. Once completed, the average of the results are used as the performance estimation. This allows for a more accurate validation to be performed on the dataset.

Once the method of determining the data to train and validate the model has been decided, the next step is to determine what metrics will be used to measure the accuracy/performance of the model. A model can be evaluated with various different methods with the most common method being the accuracy of the model. The accuracy of the model is defined as the number of correctly predicted samples divided by the total number of samples in the dataset. Table 2.5 shows a confusion matrix which can be generated for any predictive/classification models.

		Reference model	
		Event	No Event
Predicted model	Event	TP	FP
	No Event	FN	TN

Table 2.5 - Sample confusion matrix (Gromski, et al., 2009)

It shows the number of samples correctly predicted (TP; true positives or TN; true negatives) and the number of samples incorrectly predicted (FP; false positive or FN; false negative). This is representative of the outcomes of all four classes of a two-class classification model. Thus as the research problem is determining whether or not a person entered arrears, this is relevant to the research. Consequently, using the confusion matrix above, Accuracy can be defined as:

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

The inverse of the accuracy of the model is the misclassification rate, which is defined as the total number of samples incorrectly predicted divided by the total number of samples in the dataset.

$$Misclassification\ Rate = \frac{(FP + FN)}{(TP + FP + FN + TN)}$$

For example, if a dataset contained 1000 samples and 700 of these are classified correctly as either TP or TN and the remaining 300 are incorrectly classified, the

accuracy of the model would be said to be 70% while the misclassification rate would be 30%. Consequently, as the samples can only be categorised as one of two categories, 'correctly classified' or 'incorrectly classified', it can be said that the misclassification rate is the same as *1 - Accuracy*.

However using just the accuracy and misclassification rates as a measure of a models accuracy can be misleading and lead to incorrect results. This can occur when the distribution of the data is skewed as it often is in classification problems. For example in dataset mentioned above of 1000 samples, if all 700 samples predicted correctly were contained in one class and the remaining 300 were in the opposite class, the model would not have performed at a 70% rate as suggested by the accuracy and misclassification rate calculations. For these skewed datasets, it would be extremely common for the focus on the prediction rate to be based on the minority class as that would be the class of interest. Due to this, other measures of predictive power need to be utilised to give a better understanding of the model.

One of these alternative means is the Recall. Sometimes knows as the Hit Rate or the Sensitivity, the Recall is the TP divided by the combination of TP + FN. Recall is a superior measure of predictive power than just using accuracy and misclassification rate alone as it determines how well the model performed at predicting the minority class, mentioned earlier as the common focus of many classification problems. In addition to Recall, Precision is also a good measure to verify the performance of a model. Precision is the TP divided by the TP + FP. Precision informs us how often we were correct when we were predicting a positive result where-as recall informs us what quantity of actual positive results we managed to predict correctly. Specificity and False Positive rates are another two measures which can be used to analyse the performance of a model. The specificity is the measure which identifies how well the model predicted the negative class samples. The False Positive Rate is the measure to identify the number of samples identified as false positives.

$$Recall = \frac{TP}{(TP + FN)} \qquad Precision = \frac{TP}{(TP + FN)}$$

$$False\ Positive\ Rate = \frac{FP}{(FP + TN)} \qquad Specificity = \frac{TN}{(TN + FN)}$$

2.8 Conclusion

This chapter has presented an overview of the relevant literature available for the mortgages with a heavy focus on the Irish mortgage market. Clickstream data was also introduced and methods for collection of clickstream data were presented along with examples of how the data is currently used. Data mining and predictive modelling were examined and number of predictive modelling algorithms including Decision Trees, Random Forests, Regression, Neural Networks and Support Vector Machines were presented, paying close attention to how some of these models can be used for mortgage arrears classification. All of these modelling algorithms have their own strengths and weaknesses and consequently there is no one size fits all algorithm which can be used for this classification problem. Therefore a combination of these algorithms will be utilized for the purposes of this research.

This chapter also assessed various model evaluation techniques and presented various methods which can be used for model evaluation. It also put forward various reasons why certain evaluation techniques should be used over other methods. Along with this, skewed or imbalanced datasets were also discussed and various methods for alleviating the issue of imbalanced data sets were presented.

3 DESIGN / METHODOLOGY

3.1 Introduction

This chapter will provide an overview of the methodology adopted and provide an outline for the design approach of the experiments undertaken to assess whether clickstream data can be employed to improve the accuracy of a mortgage arrears prediction model. It will present the phases involved in the generation of an analytical base table (ABT) which will be used for modelling and list all of the features of the ABT along with the rationale for choosing them. Also, the software technologies used to perform exploratory data analysis and model construction will be presented along with how they were used.

Finally, the design methodology for evaluating the performance of all of the constructed models is also discussed along with reasons for chosen certain evaluation metrics over other measures.

3.2 Methodology

For this research, the CRISP-DM (Cross Industry Standard Process – Data Mining) methodology will be adopted. CRISP-DM is a data mining process approach that describes commonly used techniques to complete data mining projects. The approach is divided in six steps as per table 3.1.

Step	Description
Business Understanding	Gather business requirements and gain an understanding of the data mining project.
Data Understanding	Gather data, perform exploratory analysis to get an understanding of the data.
Data Preparation	Prepare the data for modelling. This could include handling missing data, transforming data from one format to another or data reduction amongst other tasks.
Modelling	Create models in an attempt to achieve the goals of the research.
Evaluation	Evaluate the performance of the models using defined evaluation metrics.
Deployment	Deploy the newly created model.

Table 3.1 - CRISP-DM Steps

3.3 Business & Data Understanding

The primary goal of this research is to assess whether clickstream data can be used to improve the prediction power of a mortgage arrears prediction model. To allow for this to be achieved, a mortgage arrears model must first be constructed. One key criteria which needs to be defined is how far in the future we wish to attempt to predict an account entering arrears. That is, do we want to attempt to predict if an account will enter arrears in one or two months' time for example? Due to constraints with the clickstream data which will be discussed later, it was decided to attempt to predict one month in advance. For example, if the model is attempting predicting the status of an account for 29th Sept 2017 (last working day of the month), it will be trained on a dataset ranging from March 31st – August 31st. End of months are chosen as these are the dates Lender A uses to build their periodic tables and consequently are able to be leveraged to perform projects such as this research or for other interests such as trend analysis.

Within Lender A's organisation, the arrears support unit (ASU) are the team who liaise with customers whose mortgages have entered arrears. This is carried out via various channels including phone calls and letters. Therefore if a customer was to be contacted by the ASU, there is a cost associated with it which Lender A need to cover. In addition to this cost, it may also lead to their customers being offended if they were contacted under the premise they may enter arrears which could damage Lender A's public image. Based on this, a model which has a high false positive rate (FPR) would negatively impact on Lender A's financial outcome and therefore false positives should be minimised in any model built. Consequently recall will be used as the accuracy measure for any model as it has the ability to inform what quantity of positive result was correctly predicted. In addition to this, the FPR will also be used to assess the performance of any model constructed.

The source of the data to be used in this research is Lender A's enterprise data warehouse (EDW). This EDW is a Teradata based EDW and it contains data dating back to 1998. The data contained in the EDW is an accumulation of various different source systems. For example all of the mortgage data is delivered into the EDW from the Mortgage Application System (MAS). Lender A's EDW captures and stores an historical version of various different categories of data ranging from customer personal details, through to their transactional history, all the way down to the number of heifers a farmer may have.

As of September 2017, there was approximately 81,176 active PDH mortgages on Lender A's books, with 3399 of these in arrears. This equates to approximately 4.2% of Lender A's PDH mortgages being in arrears. However only approximately 850 of these accounts in arrears have valid clickstream data. One of the reasons for this is that Lender A only began to capture clickstream data in mid-2015. Consequently, skewed datasets are certainly going to be an issue during the construction of the datasets for modelling. Therefore various different sampling techniques which were discussed earlier will be employed to enable the performance of any models to be accurately assessed.

The EDW primary key of this research is HM_LENDING_APPL_NO. This is a unique key which identifies individual mortgage accounts. This key can be joined to various other tables in Lender A's EDW to determine the customers associated with the account as well as information about the property against which the mortgage is secured. As mortgages can have one or many customers associated with them, any data collated at a customer level (I.E. transactional data) will need to be aggregated to a mortgage account level. To allow for this, a base table containing the relationship between mortgage accounts and customers will be created. Figure 3.1 shows that two customer mortgages are by far the most common type of mortgage, with 52,194 accounts falling into this category. This is hardly surprising considering it would be relatively common for couples to seek mortgages together. However there is also a large number of mortgage accounts held by a single customer, with 27,905 falling into this cohort.

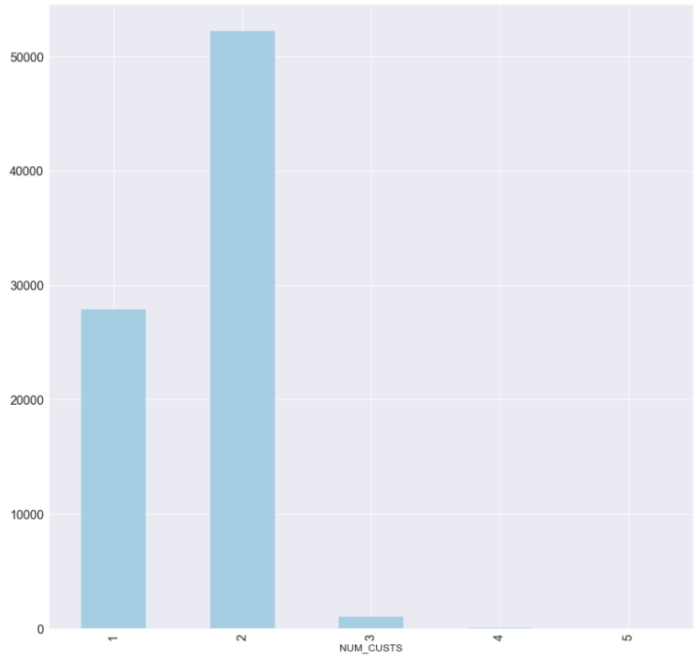

Figure 3.1 - Number of mortgages grouped by number of customers associated with the account.

Careful consideration needs to be given to specific demographical features such as a customer's age or their credit score. Aggregated values will need to be derived for each of these demographical data points. For example, for a customer's age, the average age of the customers related to a mortgage account could be used and for credit grade, the minimum (or maximum) credit card for a customer on the account could be used.

3.4 Data Preparation

Before it can be determined whether the use of clickstream features can be used as a good predictor of mortgage arrears, a base dataset and models need to be constructed to first determine the accuracy of using traditional features. This data would consist of a customer's demographical data, their transactional history, other outstanding debt (also owned by Lender A), the customer's levels of savings and finally information about the mortgage and the property the mortgage is secured against. To achieve this, traditional features will be leveraged to construct a suitable dataset and models will be constructed using this data as input to determine their validity. The customers transactional and savings history will be examined and collated for six month prior to the arrears event occurring. If a customer is not in arrears, this date range will range from 1st March 2017 through to August 30th 2017 (six months). This date range was chosen as the base date range as it avoids the unpredictable spending of the holiday period of Christmas and also encompasses a time which may be difficult for parents which is children returning to school. As mentioned earlier, a household with children has a 5% higher chance of entering arrears on their mortgage.

3.4.1 Customer Demographical Data

As shown in figure 3.1 it is very common for a mortgage account to have more than one customer associated with it and therefore when determining which demographical features to use for the modelling purpose there needs to be consideration as to how the feature can be aggregated to a mortgage account level. If an account has two or more customers, taking these customers features at their face value could lead to various highly correlated features, such as if two customers on an account had the same credit grade. There is only six mortgage accounts which have five customers associated with them and therefore if all individual customers features were to be used it would lead to a the number of demographical features increasing by a factor of five, despite the fact over 98% of the samples would have no value for customers three, four or five.

One crucial demographical feature which should be used is the income amount associated with the mortgage account. While Lender A does store the INC_AMT (I.E. salary) of its customers when it can, this data may not be correct as this data is only updated when a customer provides updated information to Lender A, for example if they apply for a loan. This is captured in both a daily and periodic (monthly, last working day of the month) fashion. Lender A have a legal obligation to ensure their data is kept up to date wherever possible. If the INC_AMT is not available, there is little which can be done to determine an accurate INC_AMT of a mortgage (I.E the salary of the mortgage account customers) as it very difficult to determine a customer's salary payment based solely their transactions.

Consequently this raises the issue of how to determine a mortgage accounts income amount. In an attempt to alleviate this potential issue, instead of using the INC_AMT value, the value of change of income will be used instead. That is, the percent of change from one month to the next for the amount of income of a mortgage account. Lender A's EDW contains a table which tracks the largest incoming transaction over the course of a month for a customer's current account. This was built with this purpose in mind.

Therefore both the level of change of this value and INC_AMT will be used. The amount which is being tracked, in both cases of INC_AMT value or transactional amount value, will be the sum of all customers associated with the account. If the value is the same from one month to the next, the value of the feature in the model will be 0 for the change between the months. If the percent of change is greater 100% (or less the -100%), the data will value will be set to 1.00 (or -1.00) to keep the data consistent with the rest of the normalised numerical values as per normalisation methods in specified subsection *3.4.5*.

Based on this, the below list of demographical features were chosen from the data available in Lender A's EDW to be used as input features for the base model.

Feature Name	Description
CURR_MAX_CR_GRADE	The maximum credit grade of all customers associated with the mortgage account (I.E. only one value, not a maximum change per customer)
NUM_CUSTS	Total number of customers associated with the mortgage account.

NUM_ACS	Total number of accounts associated with a mortgage. This is a total number of all account (current, loan etc.) associated with all customers associated with the mortgage.
MAX_CR_GRADE_CHANGE	The maximum amount of change from one credit grade to another of all customers associated with the mortgage account (I.E. only one value, not a maximum change per customer).
FIRST_TIME_BUYER_IND	Indicator of whether there is a first time buyer associated with the mortgage account.
DEPENDENT_CHILDREN_IND	Indicate whether there is dependent children associated with the mortgage account.
NORMALISE_AVG_CUST_AGE	The average age of customers associated with a mortgage account.
MTH_5_TRANS_AMT_CHG	The percent of change from month 5 to month 6 based on the largest (aggregated) deposits into the customers current account.
MTH_4_TRANS_AMT_CHG	Same as above, but for month 4 to month 5.
MTH_3_TRANS_AMT_CHG	Same as above, but for month 3 to month 4.
MTH_2_TRANS_AMT_CHG	Same as above, but for month 2 to month 3.
MTH_1_TRANS_AMT_CHG	Same as above, but for month 1 to month 2.
MTH_5_SAL_CHG	The percent of change from month 5 to month 6 based on INC_AMT change.
MTH_4_SAL_CHG	Same as above, but for month 4 to month 5.
MTH_3_SAL_CHG	Same as above, but for month 3 to month 4.
MTH_2_SAL_CHG	Same as above, but for month 2 to month 3.
MTH_1_SAL_CHG	Same as above, but for month 1 to month 2.
ACTIVE_CURR_AC	Indicator as to whether there is an active current account associated with the mortgage account.
NORMALISE_TOTAL_UNPAIDS_LAST_6_MTHS	Total number of missed payments across all accounts for the previous six months.
UNPAIDS_IN_PAST_MTH_IND	Indicator to show whether there is a missed payment associated with the mortgage from the previous month.

Table 3.2 - Customer Demographic Features

3.4.2 Transactional, Savings & Other Debt Data

Transactional data are perhaps some of the most obvious features which would need to be included in the list of features to determine the base model accuracy for this research as these features can be used to determine a customer's spending history. Similar to demographical data, transactional data will also need to be aggregated to a mortgage account level. However this should be relative straight forward as it will essentially be summing numerous transactions to a mortgage account level.

Lender A have various EDW tables which capture their customer's transactions. These tables contain data from different sources of transactions:

- Non-Card Transactions (direct debits etc.)
- Laser Card Transactions (Pre Visa debit card)
- Visa Debit Card Transactions (Standard debit card transactions)
- Credit Card Transactions (Standard credit card transactions)

All relevant transactions will be retrieved from the above sources for all relevant dates (I.E. the six months of interest in for each mortgage account). The six months of transaction spending will be averaged out to determine the average monthly spend of a mortgage account. In addition to the average monthly spend, the specific average spend, over limit percent and utilised percent of credit card will also be used as a feature. This is to attempt to cater for individuals who use credit cards to live beyond their means.

In addition to spending, the level of savings and total amount of debt associated with the account will also be utilised. Finally, the rate of repayment for the mortgage account will also be included. All of the above features are required to be normalised to ensure all data is on the same scale for modelling as discussed in subsection *3.4.5*. The complete list of features can be seen in table 3.3.

Feature Name	Description
NORMALISE_AVG_CC_AMT	Average credit card spend over the six month period.
AVG_CC_USED_PCT	The average monthly utilised percent of a credit card (I.E. amount of credit card limit used).

AVG_CC_OVERLIMIT_PCT	The average monthly over limit percent of a credit card (I.E. amount used over available credit card limit).
NORMALISE_AVG_MTH_SPEND	The average monthly spend associated with a mortgage account.
NORMALISE_TOTAL_SAV	Total savings associated with a mortgage account.
NORMALISE_REPAYMENTS	The monthly repayment amount of the mortgage account.
NORMALISE_SAV_BY_REPAYMENTS	Number of repayments the savings of a mortgage account could cover
NORMALISE_TOT_DEBT_AMT	Total debt associated with the mortgage account.
NORMALISE_TOTAL_DEBT_BY_REPAYMENTS	Number of repayments it would take to clear total other debt

Table 3.3 - Transactional, Saving and Other Debt Features

3.4.3 Mortgage Data

At the time of application for a mortgage, Lender A captures various pieces of information relating to the property against which the mortgage is to be secured. However this information is commonly not available (for example, a customer may not know the builder of their property). Consequently, many of these fields are blank for a majority of the customers. However, other fields such as the number of bedrooms in a property is populated in vast majority of cases and therefore could be used as a feature in the modelling process. In addition to information about the specific property, there is also various pieces of information captured at time of application such as the date of application, the channel through which the customer received the mortgage and the type of mortgage (such as variable rate, tracker etc.). From this, the below list of features will be used for modelling.

Feature Name	Description
APPL_YEAR	The application year of the mortgage.
MORTGAGE_PRC	The type of mortgage (I.E. variable, tracker, fixed etc.).
CHNL	The channel through which the mortgage was acquired.

HSE_TYP	The type of house (I.E. Semi-detached, detached etc.).
NORMALISE_BEDROOM_CNT	Number of bedrooms in the property
NORMALISE_NUM_YRS_OP	Number of years the mortgage has been open.
NORMALISE_NUM_YRS_LE	Number of year left on the mortgage account.
NORMALISE_DAYS_SINCE_LAST_ARREARS	The number of days since the last arrears associated with the mortgage.
CURRENT_LTV	Current Loan-To-Value of the mortgage.
NORMALISE_CURR_BAL_AMT	The current outstanding balance of the mortgage account.

Table 3.4 - Mortgage Data Features

3.4.4 Clickstream Data

As mentioned earlier, Lender A have been collecting clickstream data since mid-2015 when they deployed client side clickstream data collection onto their website with the intention of attempting to understand how their customers are interacting with their digital channel to enable them to develop a more user friendly website and ultimately attempt to sell more products to their customers. Over time, this monitoring was integrated into additional digital channels such as their mobile and tablet apps, their in-branch kiosks and into a "OneView" system which is used by staff in branch during interactions with customers.

This data are captured using various technologies across all of Lender A's digital channels and are stored as JSON (JavaScript Object Notation) logs on the systems which capture the data. From here the logs are processed through a MapReduce procedure and loaded into an Hadoop cluster. All data is loaded into Hadoop, however only authenticated data is transformed and loaded via an ETL (Extract, Transform, Load) utility into the EDW. Authenticated data are logs which can be linked back to a customer via their logon details. Non-authenticated data are not loaded into Teradata as it cannot be linked back to a customer and therefore cannot be used to drive any analytical insights when used in conjunction with other tables in the EDW. As with all data thus far, the clickstream data will need to be aggregated to a mortgage level as it is captured at a customer level.

The clickstream tables are divided into three different categories:

- Interaction
    - This table stores the interaction a customer had with a digital channel. Items such as the interaction start and end time and device used to log on are captured.
- Events
    - These are specific events a customer completed during the interaction. Items such as a loan application or a mobile phone top up are captured here.
- Attributes
    - This table contains attributes related to specific events. Items such as the loan amount for a loan application or a mobile phone number of a top up would be captured here.

Only the interaction table stores information about which customer the interaction is related to. However there is a parent-child relationship between these three tables with Interaction being the parent of Event which in turn is the parent of the Attribute table. Therefore, all events in the Event table should have a parent Interaction, and all attributes in the Attribute table should have a parent in the Event table. Due to this, all interactions, events and attributes in the EDW can be linked back to a customer.

There are approximately 1011 distinct events associated with customers which are associated to mortgage accounts in arrears. However, many of these of are an insignificant amount (I.E. one or two occurrences of specific events across the entire dataset) and are also relatively useless for modelling as they are events such as a person viewed the landing page after they logged on, since anyone who had a successful logon will see this page and consequently these events would not be beneficial to be used for this research. Therefore the list of events which are either high in frequency or which are linked to items such as loan applications will be utilised for the purposes of this research. In addition to these events, the number of logons per device and the total time a customer spent logged on to their internet banking will be used. The complete list of events along with the other clickstream features and their descriptions can be seen in table 3.5.

Feature Name	Description
NORMALISE_KIOSK_LOGON_CNT	Total number of logons from an in branch Kiosk.
NORMALISE_KIOSK_STAFF_LOGON_CNT	Total number of logons from an in branch Kiosk – Staff account.
NORMALISE_TABLET_LOGON_CNT	Total number of logons from the tablet banking app.
NORMALISE_TABLET_STAFF_LOGON_CNT	Total number of logons from the tablet banking app – Staff Account.
NORMALISE_WEB_LOGON_CNT	Total number of logons from the main website.
NORMALISE_ONEVIEW_LOGON_CNT	Total number of logons from the OneView application in branch.
NORMALISE_ANDROID_LOGON_CNT	Total number of logons from the Android mobile app.
NORMALISE_IPHONE_LOGON_CNT	Total number of logons from the iPhone mobile app.
NORMALISE_NIGHT_LOGON_IN_MINS	Total number of minutes spent logged on between hours of 00:00 – 07:59.
NORMALISE_DAY_LOGON_IN_MINS	Total number of minutes spent logged on between hours of 08:00 – 16:59.
NORMALISE_EVENING_LOGON_IN_MINS	Total number of minutes spent logged on between hours of 17:00 – 23:59.
OWN_ACC_TRANS_IND	Indictor to show whether a customer transferred funds between their own accounts.
OTH_ROI_ACC_TRANS_IND	Indictor to show whether a customer transferred funds to another Irish bank account.
SRCHED_HIST_TRANS_IND	Indictor to show whether a customer searched their historical transactions.
VIEWED_ACTI_DIR_DEB_IND	Indictor to show whether a customer viewed their active direct debits.
PAID_BILL_IND	Indictor to show whether a customer paid a bill.

VIEWED_STANDING_ORD_IND	Indictor to show whether a customer viewed their active standing orders.
PHONE_TOP_UP_IND	Indictor to show whether a customer topped up a mobile phone.
APPLIED_PER_L_TOPUP_IND	Indictor to show whether a customer applied for a personal loan top-up.
APPLIED_PER_L_IND	Indictor to show whether a customer applied for a personal loan.
CAN_DIR_DEB_IND	Indictor to show whether a customer cancel a direct debit.
VISIT_HELP_CENTRE_IND	Indictor to show whether a customer visited the help centre for internet banking.
VIEW_SPENDING_IND	Indictor to show whether a customer examined their spending history.
INTEREST_IN_PROD_IND	Indictor to show whether a customer expressed interested in products offered online (I.E. visited “our products” section of the website).
INTL_PAYMENT_IND	Indictor to show whether a customer completed an international bank transfer via internet banking.
STOP_PAPER_STAT_IND	Indictor to show whether a customer opted to stop receiving paper statements.
OPEN_STANDING_ORD_IND	Indictor to show whether a customer opened a standing order.
CAN_STANDING_ORD_IND	Indictor to show whether a customer cancelled a standing order.
DELETE_PAYEE_IND	Indictor to show whether a customer deleted a PAYEE account.
APPLY_CC_IND	Indictor to show whether a customer applied for a credit card.
ADD_PAYEE_IND	Indictor to show whether a customer applied added a PAYEE.
DEL_ACC_FROM_INT_BANK_IND	Indictor to show whether a customer deleted an account.

APPLID_FOR_OVERDRAFT_IND	Indictor to show whether a customer applied for an overdraft.
APPLIED_FOR_CC_LIMIT_INC_IND	Indictor to show whether a customer applied for a credit card limit increase.

Table 3.5 - Clickstream Data Features

3.4.5 ABT Generation

From here, the next phase is to collate all of the above features into a single ABT which can be used as input to the modelling process. The ABT will compromise of all of the above features and two other columns. The first is ARREARS_IND which will signify whether an account is in arrears with a 0 or 1, with the second column being the HM_LENDING_APPL_NO of the mortgage account. This is the key field which will allow all customer and transactional features to be aggregated to the mortgage account level. This is also the field which is used as the primary key throughout the entire data generation process.

ARREARS_IND will be populated with a 1 if a mortgage account has a "DAYS_PAST_DUE" value of greater than 10. That is, a mortgage is determined to be in arrears if a payment has not been made within 10 days of payment due date. While it was defined earlier in subsection *2.2.3* that a mortgage arrears event occurs when a payment is not made by a defined date, this does not allow for common issues which may occur and be out of the control of the customer. For example, if a separate financial organisation was processing salary payments and ran into issues with the processing, the customer may not receive their salary and consequently may enter arrears. However this is entirely out of the control of the customer and thus they should not be penalised for this. Therefore it was decided to use the 10 day figure as it gives a wide "safety net" to cater for both the scenario above and various other scenarios which may be out of the control of the customer.

The ABT contains a total number of 81,176 samples with 77,777 of these accounts not in arrears and 3,399 in arrears which represents 4.2% of accounts being in arrears. This is a significantly skewed dataset and therefore careful consider is needed to ensure the accuracy of any modelling process is not positively or negatively impacted by it. This will be discussed further in subsection *3.5*.

Many of the features contained within the ABT are numerical features which can vary widely in their value. Also, not all numeric features are represented in the same units.

For example, the value of the clickstream logon times are in minutes while the logon counts are simply counts. Therefore if all numerical features were used in the model without any transformation, the results would vary widely as the values are inherently different both in terms of size and what they are measuring. Due to this, it is required to normalise all numerical features to the same scale. To achieve this, all features will be normalised to a value between (0 – 1) or for negative features such as AVG_MTH_SPEND, (-1 - 0) as the source values are negative values. The formula to calculate the normalised values is below. For the features which are required to be normalised to between (-1 - 0), the results of this equation is multiplied by -1.

$$\frac{x - \min(x)}{\min(x) - \max(x)}$$

Where *x* is the value we wish to normalise, *min(x)* is the minimum value of the entire range of *x*, and *max(x)* is the maximum of the entire range of *x*.

Once all of the numerical features have been normalised to the same scale, they were ready to be used as input to the modelling process. However the ABT still contained some categorical features such as MORTGAGE_PRICING and HOUSE_TYPE. As many machine learning algorithms perform mathematical calculations using the input features, they cannot process non-numerical data. Therefore to allow these features be used in the models they needed to be converted to dummy variables. This is a process where-by numerous new features are created, one for each category contained within the original categorical feature. Each of these features will contain a 0 or 1 to represent the value of the category. Figure 3.2 shows the HOUSE_TYPE feature being translated to a set of dummy variables.

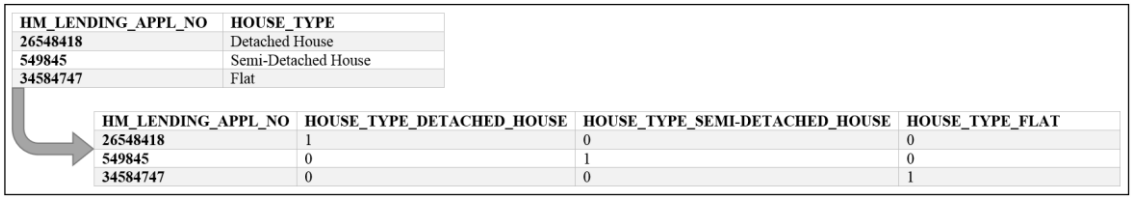

Figure 3.2 - Dummy Variable Creation Example

This process was not carried out during the generation of the ABT as it would have led to the number of columns on the table increased significantly. As can be seen in figure 3.2, the HOUSE_TYPE column was translated to three new columns. Therefore, this process would be carried out during the modelling phase where the extra columns could be created in memory and not be required to be permanently stored in the EDW.

3.4.6 Software Used

As already mentioned, the source of all of the data used for this research was Lender A's Teradata EDW environment. SQL was used to collate all of the many sources tables into the final ABT discussed above. From here, an ETL tool was used to extract the data and create CSV (Comma Separated Values) files. It was decided to use an ETL tool for this process to avoid issues with manually copying data from SQL query results. As the EDW was the source of this data, it was relatively clean at the time of ABT creation with minimal data quality issues.

Once the data had been extracted and collated into CSV files, Python was used to perform data analysis and modelling. Python is a free and open-source programming language developed by Guido van Rossum and released in 1991 (Rossum, 2009). While Python was not specifically designed to perform data mining related activities, its ease of use and code readability led to it becoming a popular tool in the data mining community. Within Python, there are various packages which can be utilised to perform data mining activities. For this research, the packages used were SKLearn (AKA Sci-Kit Learn), NumPy, Pandas and MatPlotLib. SKLearn is a machine learning package for Python developed to allow for machine learning to be carried out with Python. It features many classification algorithms such as SVM, Random Forests, Decision Tress, K Nearest Neighbour and Adaboost amongst others as well as regression and clustering algorithms. NumPy is a package which allows for scientific computing within Python. It's allows for the creation and processing of n-dimensional array objects and also can be used to generate random numbers. Pandas is a package which was developed to enable data manipulation and analysis within Python. It allows for CSV files to be read into memory to allow for the data to be processed both by itself and other packages such as SKLearn. Finally, MatPlotLib is a package which enables Python to create 2D visualisations of the data it is working with. It can produce bar charts, histograms, pie charts and scatter plots as well as various complex visualisations. The majority of the visualisations for this research were created in Python using both the MatPlotLib package and Pandas default plot function. The Jupyter Notebook was used as the development environment for this research.

In addition to Python, a graphical utility was initially used to enable rapid analysis of the source dataset. This tool is called Knime (pronounced Nime) and is a free to use graphical analysis and modelling tool. It functions by utilising "nodes" which are pre-

built utilities and algorithms and are connected in a process flow type manner. Knime allows for datasets to be read in, and models constructed very quickly without having to write any code. Compared to Python, Knime has the advantage of being a graphical tool which ensures its ease of use and enables the construction of models rapidly. Figure 3.4 show an example of a Knime workflow which is the construction of a decision tree model for the Kaggle Titanic completion. Without any feature engineering, this model scores an accuracy of 75%. However a disadvantage of Knime is the fact it only has a set list of functions which it can complete (I.E. a set list of nodes). Therefore if something was required for which there was no node available, it may be difficult to achieve that task in Knime. Knime does enable users to write Java code if they wish, however this aspect of the tool was not utilised for the purposes of this research. This aspect would one advantage Python would have over Knime. It allows for a much broader control of the modelling process and can allow for minor items to be tweaked to the researcher's preference. Also there is a far larger online community of Python users and therefore a much broader support network.

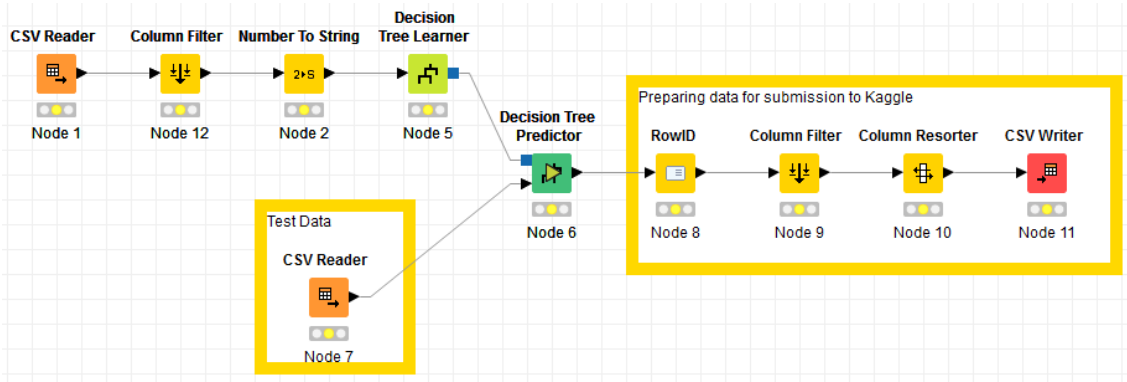

Figure 3.3 - Knime Sample Workflow

3.5 Model Construction

Once the ABT from subsection *3.4.5* had been created, the next phase of the research is to build classification models to determine if an account is at risk of entering arrears. As already mentioned, the data contained within the ABT is heavily skewed towards non-arrear accounts and therefore if the entire dataset is used for modelling it may lead to some of the issues discussed in subsection *2.7*, where-by if the model predicted every account would not be in arrears it would be correct approximately 96% of the time as approximately 4% of the dataset is in arrears. To avoid this occurring, various different subsets of data will be extracted from the entire ABT. Initially, base models will be

created using a base dataset and these models will then be applied to other datasets after which their accuracy will be assessed.

The list of datasets to be created can be seen in table 3.6. *Mortgage1.csv* will be the base dataset created and utilised for developing the models which will be used for the subsequent datasets. This dataset will only contained the traditional features. This is determine how well any model performs on the base data without clickstream data applied. This also allows a larger sample size to be used for training as we can use every account in arrears and not just the 850 in arrears which have clickstream data related to them.

The primary datasets of the research are *Mortgage2.csv* and *Mortgage3.csv*. They will be the same dataset, except *Mortgage3.csv* will have synthetic sampling applied to increase the number of accounts in arrears. They will contain all of the same traditional features as *Mortgage 1.csv*, but will also contain the clickstream features. *Mortgage4.csv* and *Mortgage5.csv* will only contain clickstream features. These will be employed to determine whether the standalone clickstream features can be used as a predictor for mortgage arrears. All of the records in the datasets with be a random sample from their relevant cohort (I.E. the 12,000 non-arrears will be a random sample of the 77,777 non-arrear accounts). The number of samples in each dataset was chosen in an attempt to give a broad as possible understanding of the data to the models. As mentioned above, the entire ABT could not be employed as it is too heavily skewed.

Dataset Name	Content	Arrears/Non-arrears split
Mortgage1.csv	Base Dataset with no clickstream features.	Arrears: 3399 Non-Arrears: 12,000
Mortgage2.csv	Same as above, with clickstream added.	Arrears: 850 Non-arrears: 4000
Mortgage3.csv	Same as Mortgage2.csv, but with extra arrears created via synthetic sampling via SMOTE	Arrears: 4000 Non-arrears:4000
Mortgage4.csv	Only contains clickstream features	Arrears: 850 Non-Arrears: 4000

Mortgage5.csv	Same as Mortgage4.csv with addition arrears created via SMOTE.	Arrears: 3500 Non-arrears:3500

Table 3.6 - Input Datasets

As discussed in chapter 2, a common issue while attempting to create a classification model is overfitting the model to its training data so as to prevent it generalising well to previously unseen data. In an attempt to avoid this occurring, consideration will be undertaken with regards to determining what data to use to train the models. Each model built will be trained using a training dataset, which will be a subset of the overall input dataset. To achieve this, the source data needs to be separated in "training" and "test" datasets. Pythons "*train_test_split*" function will be employed to create these training and test datasets. This allows the source dataset to be split based on a user defined amount. For this research, the split with be a 70/30 train/test stratified split. The stratification of this split will be based on the target features and consequently the overall split of the initial dataset will remain across the training and test subsets.

Further to this, other methods which can be employed to avoid overfitting models were also examined in subsection *2.7*, with one of the methods discussed being cross-validation. For this research, cross-validation will be employed to validate the accuracy of the models more robustly. A 10-fold cross-validation will be used, with the accuracy measure being the recall of the model. The model accuracy will be determined to be the average of all of the results from the 10-fold cross-validation. To achieve this, SKLearn's "*cross_val_predict*" will be used and the performance metrics will be calculated from the outputted confusion matrix.

Initially a base model will be built to establish a baseline accuracy against which further models can be compared. The model will be a decision tree, due to their ease of understanding. Following from this model, various different classification algorithms will be applied, tuned and tested against the primary base dataset *Mortgage1.csv* to establish which algorithms will be utilized further in the research. These models will include "traditional" algorithms such as logistic regression, SVM and k-NN. In addition to the traditional algorithms, multiple ensemble models including random forests, gradient boosting decision trees and Adaboost will also be examined. The highest performing models will then be used in all further phases which will establish the impact, if any, clickstream features will have on a mortgage arrears prediction model. Finally,

the same models will be used to determine the performance of standalone clickstream features are mortgage arrears prediction.

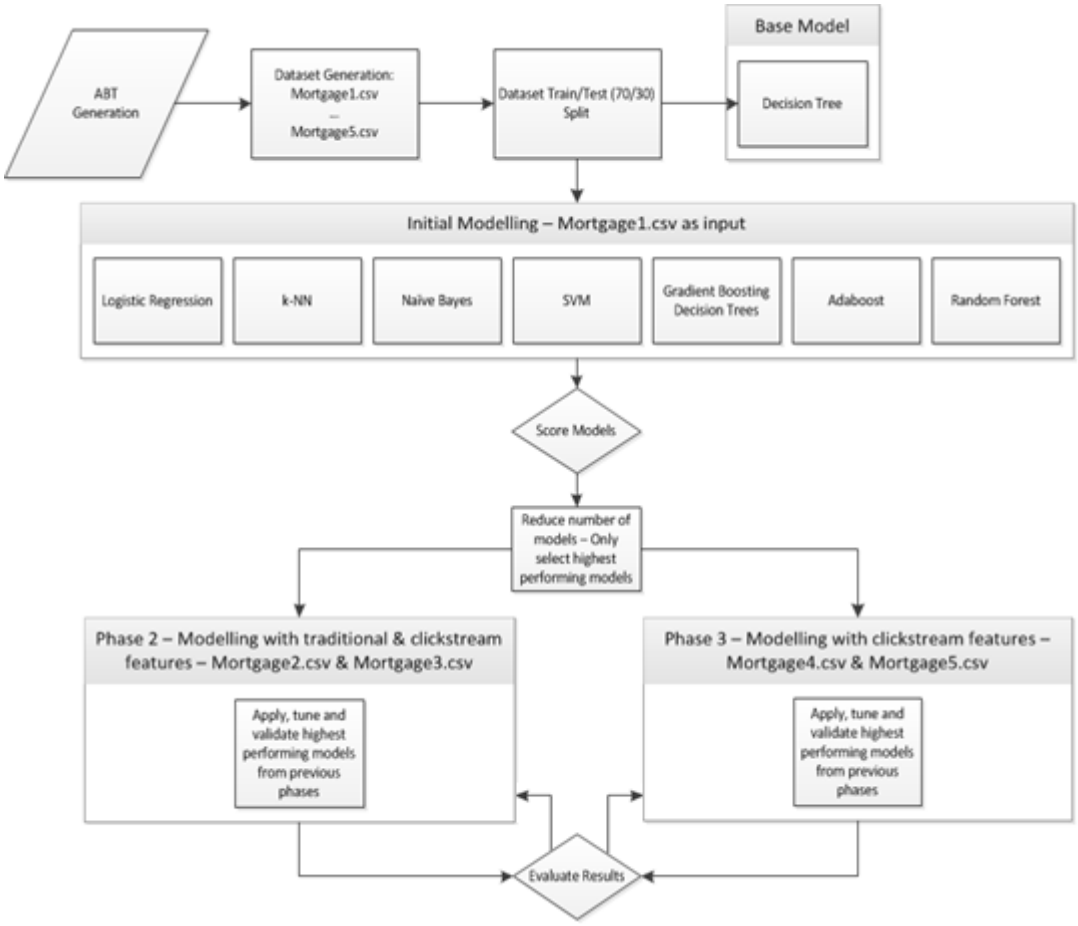

Figure 3.4 - Overview of design process

3.6 Evaluation

Measuring the performance of a model is a crucial step in any data mining project. It is during the evaluation phase where the performance of any constructed model is discovered along with determining if issues such as overfitting are occurring. Several evaluation techniques were examined in subsection *2.7* and many of these will be employed to validate the accuracy of any model produced. As the source dataset for this research is heavily skewed towards accounts not in arrears, using the commonly used *accuracy* measure (percentage of records classified correctly) measure isn't suitable for reasons already mentioned in subsection *3.5*. Consequently other metrics are required to be utilised.

Various evaluations metrics will be used to validate the accuracy of the models created for this research. An important aspect while evaluating the performance of a model is to construct a confusion matrix from output of the models predictions. From here, multiple

different evaluation metrics can be calculated including the recall, precision, specificity and the FPR. As the target feature for this research is a minority of the dataset, both recall and precisions will be used as they are a good metric to determine how well a model is performance against the minority data.

A confusion matrix will be produced for each of the models run and values for recall, precision and FPR will be calculated. From here, all of the measures will be assessed and compared to determine the highest performing model on the defined datasets. The highest performing models will be examined more closely in an attempt to improve their prediction power. However, reducing the false positives will be of high importance for any measurement and therefore the FPR will be an important metric in addition to recall and precision while tuning any models.

3.7 Conclusion

This chapter examined the design approach and methodology adopted and also presented the fundamental objective of the research. The approach to the construction of the ABT was reviewed along with a detailed analysis of the features. In depth descriptions of the clickstream data were conveyed along with reasons for choosing certain aspects of the data over other aspects for the modelling phase.

The various different software tools and packages employed for this research were discussed and their features were presented. Finally, the approach to model construction was established in addition to information as to how the models will be evaluated along with reasons for using certain evaluation metrics over other metrics.

4 IMPLEMENTATION / RESULTS

4.1 Introduction

This chapter will introduce and describe the implementation of experiments undertaken in assessing the impact clickstream data can have on a mortgage arrears prediction model. The methods of modelling are discussed along with the justification for employing certain modelling algorithms over other algorithms.

Initially a base model will be constructed. This model is a decision tree and it will be the model against which further constructed models will be compared to assess their performance. Several phases of experiments are presented along with the result output of each experiment. Finally the results will be summarized and evaluated to determine the performance of the models.

4.2 Modelling

For the purpose of a predictive modelling project, it is common to apply various different modelling algorithms to the dataset in an attempt to establish which has the highest performance measures. This phase will also identify algorithms which are performing poorly on the datasets and therefore should not be used for the purposes of the project. Consequently for the purposes of this research, various modelling algorithms will be applied, tuned and tested. From here, only the highest performing models in terms of recall accuracy and a low false positive rate (FPR) will be utilised further in the research.

As mentioned in subsection *3.4.5*, all datasets are still required to be processed further to convert categorical features to dummy features. To achieve this, Pandas *get_dummies* function was used. Figure 4.1 shows how this was implemented. A function was defined to accept the data which needed to be converted to dummy variables along with a list of columns to create dummies for. The data is processed via a for loop and returns a *DataFrame* containing the original dataset with the categorical features removed and replaced with dummy variables.

```
#Specify which columns against which to create dummies
todummy_list = ['APPL_YEAR','MORTGAGE_PRICING','CHANNEL','HOUSE_TYPE']

#Function to deal with creating all dummy variables.
def create_dummies(data, todummy_list):
    for x in todummy_list:
        dummies = pd.get_dummies(data[x], prefix=x, dummy_na=False)
        data = data.drop(x,1)
        data = pd.concat([data,dummies],axis=1)
    return data
```

Figure 4.1 - Code to create dummy variables for categorical variables

It should be noted the feature CURR_MAX_CR_GRADE is a categorical feature but does not appear in the *todummy_list* in figure 4.1 While it was originally planned to use this feature, it became apparent after dummy variable generation that the feature was heavily correlated (77%) with the target feature and therefore this feature was removed from the analysis. A correlation matrix was generated to assess the correlation across all of the features to ensure it would not lead to any concerns during the research. However this image was too large to be included here.

A small number of samples across the dataset contained missing values for 2 fields, namely NO_BEDROOM_CNT and NUM_YRS_LEFT. These missing values was replaced with the median value from the entire dataset. This replacement was completed using Pandas *fillna* function.

All of the performance metrics presented below were derived using SKLearn's *cross_val_pridict*. This performs a prediction of the test data using 10-fold cross validation against a model trained on the corresponding training data. This method was chosen over the standard *score* function of the models as it provides far more robust testing and validation of the models.

4.2.1 Traditional Features

The initial phase of this modelling process is to get an understanding of how models are performing on a base dataset before any clickstream features are applied. The primary reason for this is sample size. As there are only approx. 850 mortgage accounts in arrears which have valid clickstream data associated with them, we cannot used many non-arrear accounts as this will lead to a heavily skewed dataset which will lead to problems discussed previously. Consequently *Mortgage1.csv* will be used as the base features set. This dataset contains a much larger subset of data from the overall ABT and therefore should provide a good baseline for the purposes of this research.

Before various modelling algorithms are applied to a dataset, it is good to have a baseline model against which the results of other models can be compared. This provides a more robust evaluation method over employing the standalone *no information rate*, which is an accuracy of 50%. This is the equivalent of randomly assigning the arrears status to an account. A decision tree will be utilised for this purpose. A decision tree was chosen as they are easy to understand and visualise and can easily identify the highest grossing features of the dataset. The decision tree was built using SKLearn, employing its *Tree* and *DecisionTreeClassifier* packages. Originally the tree was built using the default parameters set out by the package which employs G*ini Impurity* as the splitting condition and continues to grow until there is one sample per leaf node. This resulted in a 70% accuracy, with a FPR of 9%. However as this tree continued to grow until there was only one sample in each leave node, it was substantial in size and likely be susceptible to overfitting. Consequently the minimum number of samples in the leaf node was set to 10 and the splitting condition was changed to information gain and the decision tree was rerun. This increased the accuracy to 71% and reduced the FPR 6%. From here, the minimum number of samples per leaf node was increased and the accuracy was determined. Based on this, the base decision tree against which further models will be assessed was determined to be a tree with a minimum number of samples of 70 and a splitting condition of information gain. Once the minimum number of samples increased past 70, the precision of the model began to drop despite the accuracy increasing.

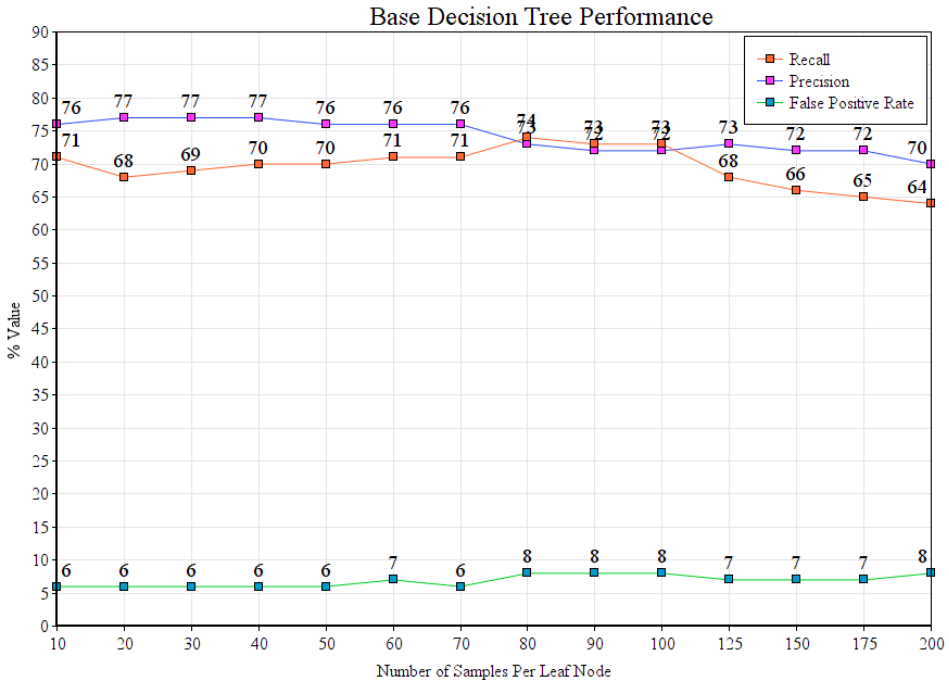

Figure 4.2 -Base Decision Tree Performance

Interestingly the tree with a minimum left node size of 10 has similar performance to the tree with 70 samples in the leaf node however this drops as the minimum number of samples increase. As the tree with a minimum number of 70 samples would be a smaller and therefore a less complex tree, it was chosen over the more complex tree with a minimum leaf node size of 10.This was for reasons discussed in subsection *2.5.1* where (Quinlan, 1986) states that out of two trees, the simpler one should be chosen.

From here, the next step was to apply various other modelling algorithms to establish their performance. Originally based on the research carried out for chapter 2, it was planned to employ SVM and Random Forests as the primary modelling algorithms. However as the research progressed, it apparent this was not a good approach as it was severely limiting the possibility of answering the goal of the research. Consequently the list of modelling algorithms in table 4.1 were used.

Algorithm	**SKLearn Package**
Random Forests (RF)	ensemble.RandomForestClassifier
Gradient Boosting Decision Trees (GBDT)	ensemble.GradientBoostingClassifier
Logistic Regression (LR)	linear_lodel.LogisticRegression
SVM	svm.SVC
K-Nearest Neighbour (k-NN)	neighbors.KNeighborsClassifier
Adaboost	ensemble.AdaBoostClassifier
Naïve Bayes (NB)	naive_bayes.GaussianNB

Table 4.1 - Modelling Algorithms Tested

All of the above algorithms were run against the same dataset and their performance was established. Rather unsurprisingly all of the ensemble modelling techniques (RF, GBDT and Adaboost) outperformed the other models. LR, SVM, k-NN and NB all performed poorly in terms of accuracy and high FPRs. Consequently, these models will not be used any further in this research. These models along with their results will be discussed further in subsection *4.2.1.1*.

The three ensemble modelling algorithms performed relatively well using their default parameters. All of the models employed decision trees in their prediction process. Therefore a key parameter to be tuned is the number of decision tress created during the training process. All of the models were run with a tree count of 5, 15, 25, 50 100, 250, 400, 500 and 1000. This range of tree count was chosen as it gives a broad range of

numbers to establish a broad spectrum of performance metrics for the models. The results can be seen in figures 4.3, 4.4 and 4.5.

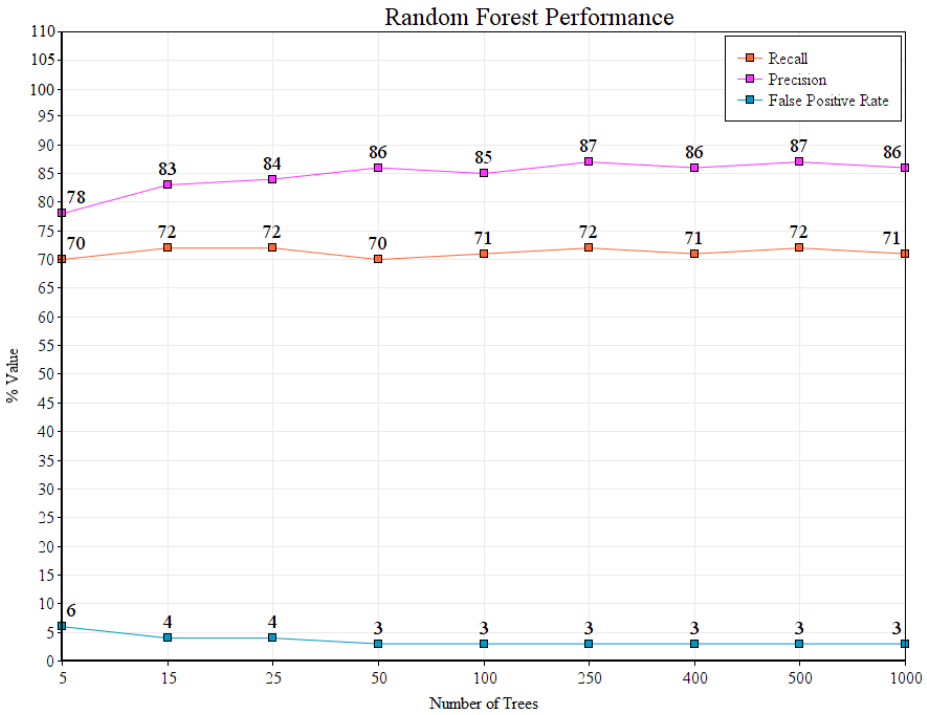

Figure 4.3 - Random Forests Performance for Mortgage1.csv

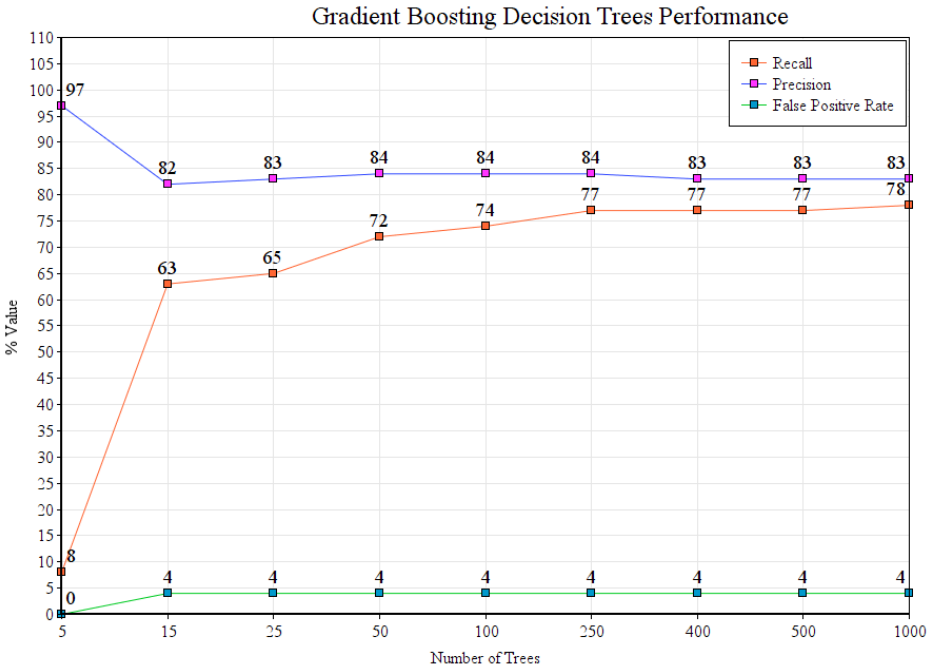

Figure 4.4 - GBDT Performance for Mortgage1.csv

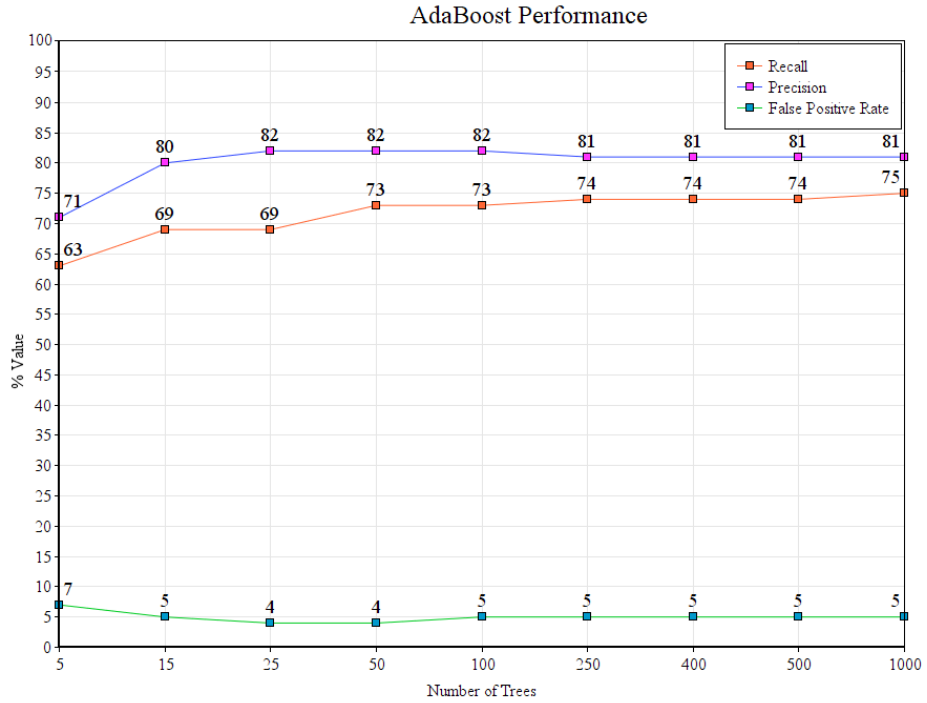

Figure 4.5 - Adaboost Performance for Mortgage1.csv

Based on these, the optimal models for the base dataset are below. Apart from the specified parameters, all other parameters are defaulted as per their respective SKLearn packages. While some models did perform slightly better than the below models, the training time on these models was also substantially higher and therefore they weren't chosen.

Random Forest: *n_estimators*=250, *criterion* ='entropy' (250 trees, split on Information Gain)

GBDT: *n_estimators*=250

Adaboost: *base_estimator*=DecisionTreeClassifier, *n_estimators*=250

4.2.1.1 Poorly Performing Models

While ensemble models performed relatively well at the prediction task, many other standard models struggled. Using recall as the accuracy measure, LR, SVM, k-NN and NB all performed poorly, with NB performing the worst overall due to its tendency to have a very high FPR. Another issue was with SVM and the amount of time it takes to train when compared to other method such as the ensemble methods which are quicker to train even when training with a forest size of 1000 trees.

LR performs poorly if all parameters are left to the default value. However if the parameters are changed, an accuracy of 78% can be achieved. This seems like an

excellent result, until the FPR is examined, which was 17%. Consequently the parameters which caused this high FPR was reverted back to the default value and only the C parameter was tuned in an attempt to increase performance. However this did not produce any favourable results, as can be seen in figure 4.6. All of the C values examined produce very similar results.

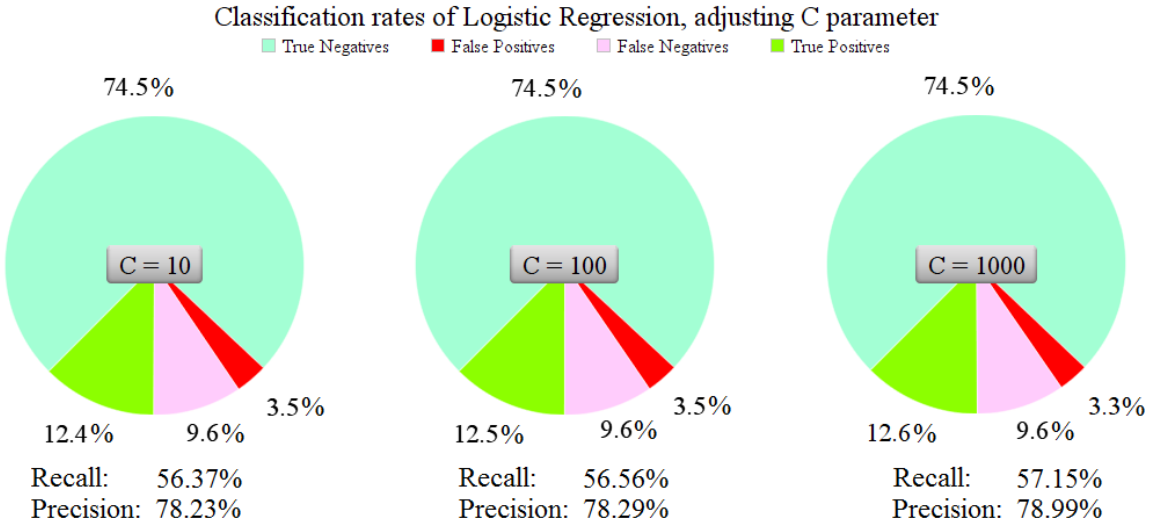

Figure 4.6 - LR results

Similar to LR, k-NN performed poorly despite parameter tuning. Using the default Euclidean measure as the distance measure, the number of neighbours was adjusted and resulted in little performance improvement, as shown in figure 4.7.

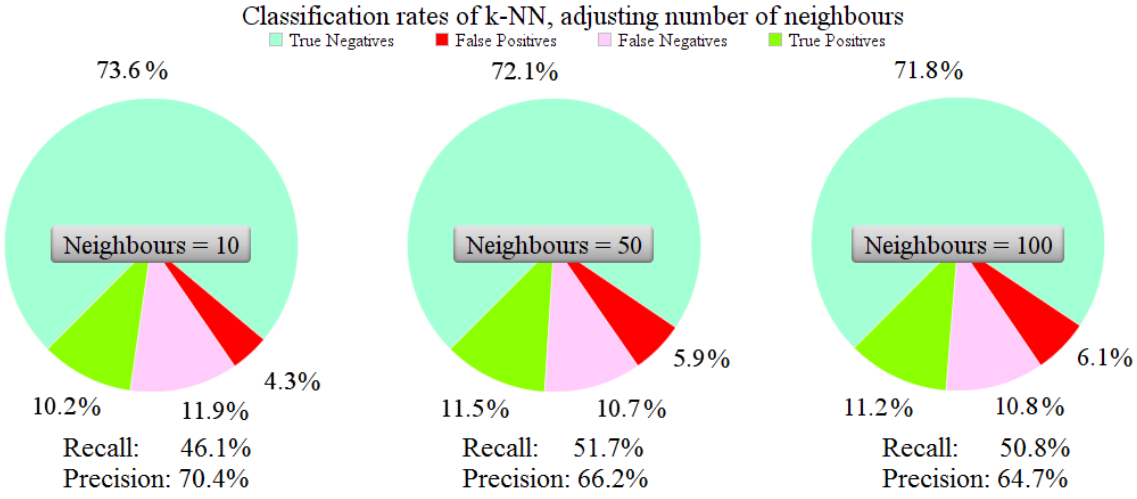

Figure 4.7 - k-NN Results

Unlike LR and k-NN, SKLearn's implementation of NB doesn't have any parameters which can be tuned and therefore if the model is performing poorly there is no way to improve the performance. While the recall of NB is considerable at 90%, the FPR is

45%, which essentially leads to it being unusable for this research. The confusion matrix for NB can be seen in table 4.1.

	Predicted Non-Arrear	**Predicted Arrear**
Actual Non-Arrear	1969	1631
Actual Arrear	94	926

Table 4.2- NB Confusion Matrix

Like the aforementioned models, SVM performed very poorly both in term of accuracy and FPR. Despite adjusting the *C* parameter, the maximum accuracy attained by SVM was 55% with a FPR quite high at 11%. This accuracy was achieved with a *C* parameter of 1 and a kernel of radial base function (RBF). When the *C* parameter was increased to 100, the accuracy dropped to 44% along with the FPR which dropped to 5%. However as this accuracy is less than the *no information rate*, it essentially means the model provides zero predictive power. Also, in addition to performing very poorly, SVM also took an extremely long time to train when compared to other models, taking up to fourteen times longer to train than LR. If the kernel was changed to be polynomial, the model would not run at all and would fail after running for approximately nineteen minutes.

4.2.2 Traditional and Clickstream Features

Once the performance of various models on traditional features had been established, the next step was to determine the models accuracy on a dataset containing the traditional features in addition to the clickstream features. This dataset was *Mortgage2.csv*. A known potential issue with this dataset is the sample size. It only contains 4852 samples, of which 852 are in arrears. These numbers reduce down even further when the data is split into training and test datasets with the 70/30 training/test split.

Consequently it may be difficult to construct a high performing model for this dataset. Therefore a second dataset containing traditional and clickstream features was created. This second dataset was *Mortgage3.csv* and this dataset is identical to *Mortgage2.csv* except it has additional accounts in arrears created via SMOTE. This dataset contains 8000 samples, an even split of accounts in arrears and not in arrears. As this dataset now contained an even split, recall was no longer required as the primary accuracy measure. Therefore standard accuracy could be employed. However this accuracy measure should not be directly compare to the recall accuracy of the other models. Only recall accuracy measures will be compared.

Knime was used to create these SMOTE records, as can be seen in figure 4.8. The data was first shuffled and then passed into a SMOTE node. From here, it was specified to use 10 of the closest neighbours, via k-NN, to create the synthetic samples. This created 3148 synthetic samples in addition to the 852 real samples to bring the total number of in arrears samples to 4000 in *Mortgage3.csv*.

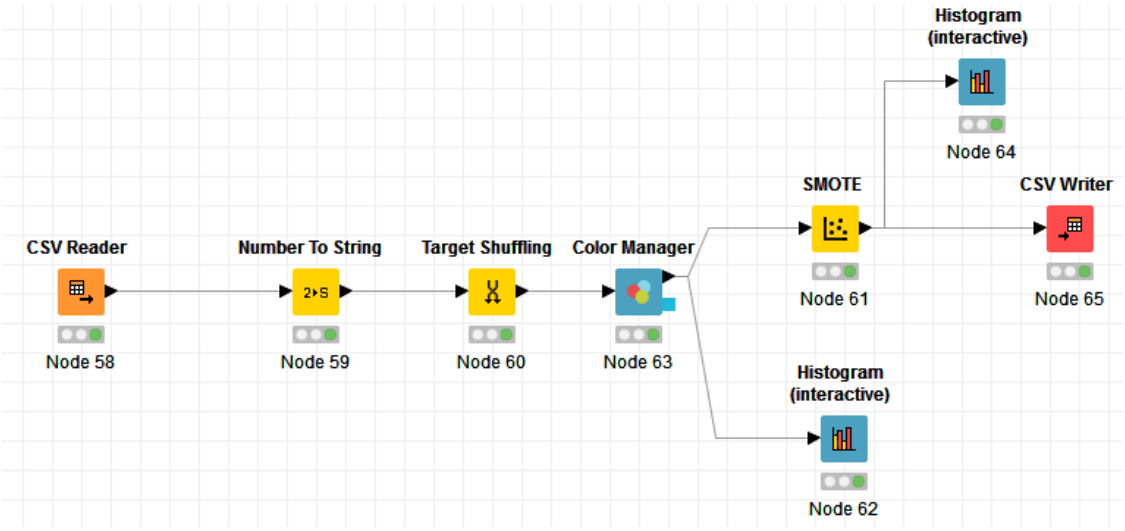

Figure 4.8 - Creating SMOTE samples via Knime

The datasets are now ready to be fed into modelling algorithms to determine the impact, if any, clickstream can have on a mortgage arrears prediction model. As determined in subsection *4.2.1*, ensemble models will only be used for this processes due to how poorly various other models applied to the base dataset which contained traditional features.

Initially models were run against these datasets with the clickstream features removed to determine the accuracy of standalone traditional features. While the results of the modelling in subsection *4.2.1* only used these traditional features also, there was far more samples on which the models were trained and consequently may perform differently.

	Mortgage2.csv			**Mortgage3.csv**		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	47.3	66.8	5	74.3	78.7	20
15	44.9	74.7	3	76	87.4	11
25	44.8	74.4	4	76.7	91	8
50	39.4	73.8	3	75.2	95.2	4
100	44.1	74.3	3	75.1	97.4	2

250	45.7	76	3	76.1	98.2	1
400	46.9	75.5	3	75.6	98.2	1
500	43.8	75.6	3	75.6	98.3	1
1000	46.5	77.8	3	75.6	98.5	1

Table 4.3 - Results of Random Forests on Mortgage2.csv and Mortgage3.csv with clickstream features removed

As was somewhat expected, the performance of RF on *Mortgage2.csv* was extremely poor, performing worse than randomly choosing the arrears status of an account. This was due to the number of samples in the dataset. However the same models performed well on *Mortgage3.csv*, which is the same as *Mortgage2.csv* with additional in arrears accounts generated via SMOTE in Knime. A RF with 250 trees was the highest performer on *Mortgage3.csv*, as it was on *Mortgage1.csv* in subsection *4.2.1*. Similarly with GBDT, the highest performing model was 250 trees. Once the number of trees increased passed 250 for GBDT, the FPR grew as can been seen in table A.1 in appendix A. However in contrast to RF, GBDT performed relatively well on the *Mortgage2.csv* dataset which was unexpected (achieving a 67% recall and 5% FPR with 250 trees). Finally, Adaboost looked to be performing well, but once the number of trees passed 50, the FPR began to increase significantly, achieving a 13% FPR with 250 trees for *Mortgage3.csv*. This would not be an acceptable rate for reasons described in subsection *3.3*. For *Mortgage2.csv*, like GBDT, Adaboost performed well achieving an accuracy of 65.2% and a FPR of 5% with 50 trees. Similarly with *Mortgage3.csv*, the highest performing Adaboost model was 50 trees which performed with a recall accuracy of 70% and a FPR of 7% (full results are attached in appendix A).

These models were then rerun with the clickstream features added back into *Mortgage2.csv* and *Mortgage3.csv* respectively. With these additional features added, RF performed relatively well compared to the dataset without the features for *Mortgage2.csv*. Similar to the results from table 4.3, a RF with 250 trees was a good performer, along with all models with a higher tree count. The addition of the clickstream features did not lead to any significant change in the model performance for *Mortgage3.csv*, but did for *Mortgage2.csv*. However, the results for Mortgage2.csv were still only marginally better than the no information rate. Without clickstream features, a RF or 250 trees achieved a recall of 76.1% with a FPR of 1%, while a model with clickstream features and 250 trees achieved a recall of 74.9% with a FPR of 1%, which

are very similar results. Also, the precision was slightly higher with the clickstream features included.

	Mortgage2.csv			Mortgage3.csv		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	48.8	68.7	5	74	77.3	21.6
15	49.6	72.2	4	76	87.3	11
25	54.7	74.9	4	75.4	90.1	8
50	50.8	79.8	3	74.1	94.4	4
100	55.9	79	3	74.7	96.4	3
250	52	77.8	3	74.9	98.7	1
400	53.5	80.6	3	75.3	98.2	1
500	53.1	77.7	2	75.5	98.1	1
1000	53.9	80.7	3	75.3	98.4	1

Table 4.4 - Results of Random Forests on Mortgage2.csv and Mortgage3.csv with clickstream features included

GBDT produced similar results, with 250 tress achieving the highest performance overall with a recall of 73.3% and a false positive rate of 5%. Once the number of trees increased past 250, the FPR grew steadily. For *Mortgage3.csv*, the addition of clickstream features provided little change to the accuracy of the GBDT models, as can be seen in figure 4.9. However for *Mortgage2.csv*, the addition of clickstream features increased recall accuracy by approximately 3% while maintaining the same FPR, which can be seen in figure 4.10. The full list of results can be seen in table A.2 in appendix A.

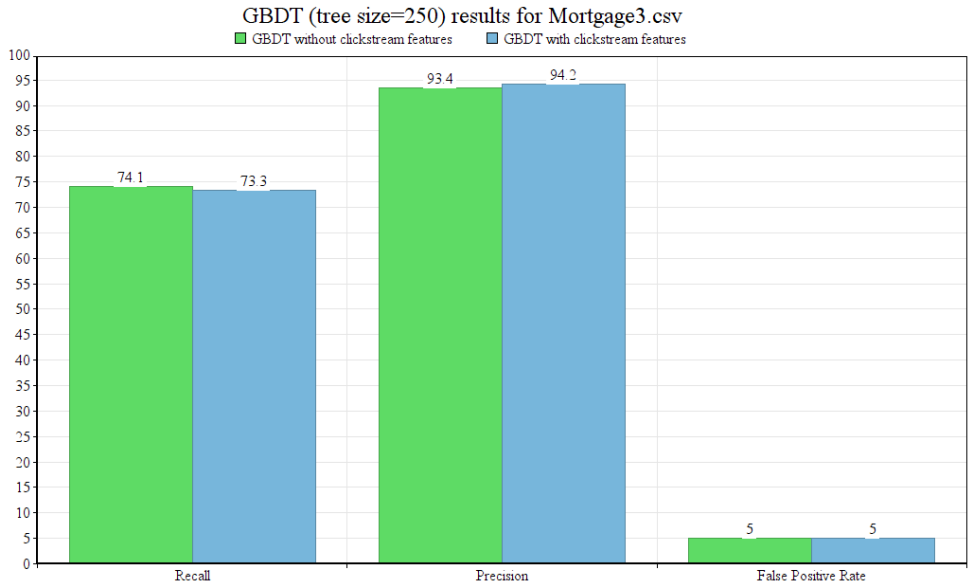

Figure 4.9- GDBT for Mortgage3.csv

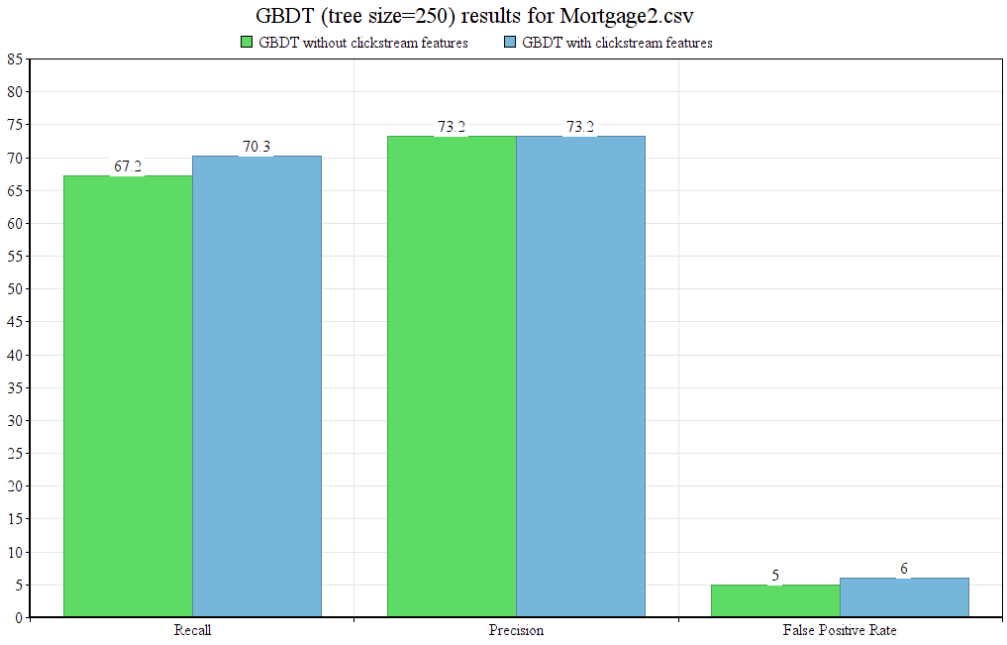

Figure 4.10 - GBDT for Mortgage2.csv

As with the other ensemble models, Adaboost seemed to be achieving moderate results but once the number of trees surpassed 50, the FPR grew steadily, as observed in figure 4.11. For *Mortgage2.csv*, at 50 trees, Adaboost had a performance of 65.2% without clickstream features compared to 62.9% with clickstream features. The results were broadly similar for *Mortgage3.csv* where the accuracy was slightly decreased from 70% to 68.8% when the clickstream features were added. The full list of results can be seen in tables A.3 and A.4 in appendix A.

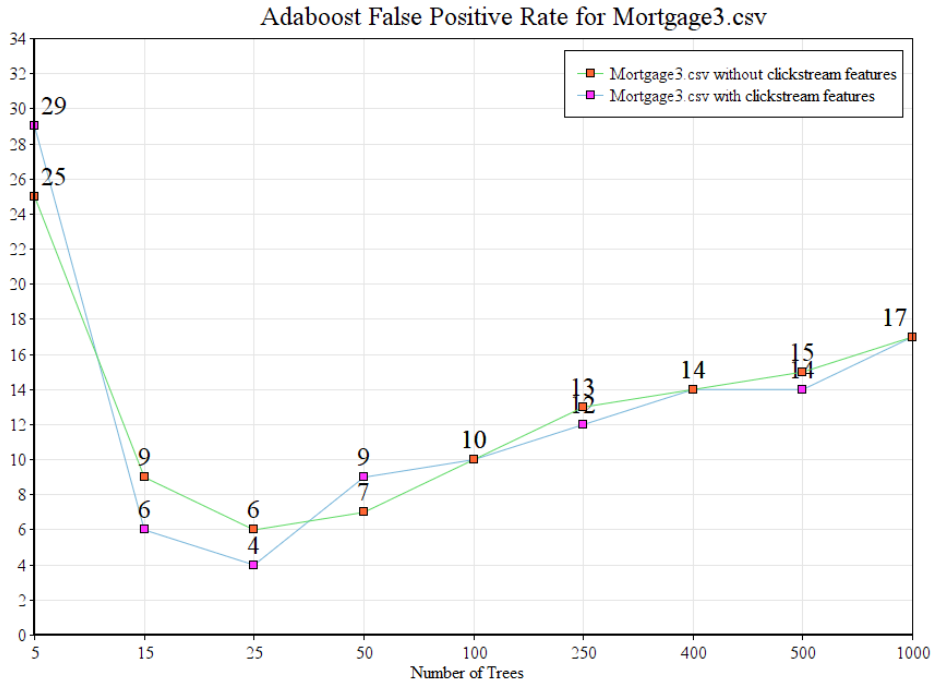

Figure 4.11 - Adaboost FPR for Mortgage3.csv

4.2.3 Clickstream Only Features

The final portion of this research is to establish the performance of models when run against standalone clickstream features. The same models used until to this point for traditional and clickstream features will be reused with all traditional features removed from the dataset. *Mortgage4.csv* and *Mortgage5.csv* were created for this purpose. Similar to *Mortgage2.csv* and *Mortgage3.csv* above, *Mortgage5.csv* is the same dataset as *Mortgage4.csv* except it has additional in arrears account added synthetically via SMOTE.

As *Mortgage4.csv* has the same samples sizes of *Mortgage2.csv*, it has the same issues with regards to sample sizes preventing the creation of high performing models. As expected, RF performed very poorly on *Mortgage4.csv* achieving a maximum recall accuracy of 17%, which is again worst the assigning an arrears status at random. However as with the results for *Mortgage3.csv*, the highest accuracy for a RF for *Mortgage5.csv* was obtained with a 250 tree RF, with a recall accuracy of 72% and a FPR of 15%. While this FPR is substantially higher than all previous RF models, it is relatively low when compared to the rest of the RF constructed for this portion of the research. However, this is still a very poor model for the goals of this research as the FPR is very high. Table 4.5 contains the entire result set for RF against clickstream only features.

	Mortgage4.csv			**Mortgage5.csv**		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	17.2	32.1	9	70.8	77.6	20
15	14.1	39.1	5	71.4	78.2	20
25	12.9	45.2	4	71.3	79.2	19
50	12.5	42.1	4	71.6	80	18
100	12.1	41.9	4	70.9	81.1	16
250	12.9	45.8	4	72.1	81.9	15
400	12.5	44.4	4	71.8	82	16
500	13.3	45.6	4	71.3	82.1	16
1000	12.5	45.1	4	71.7	81.7	16

Table 4.5- RF results for clickstream only datasets

Similarly GDBT performed abysmally on *Mortgage4.csv*. This was surprising considering the results presented earlier for GBDT when it was applied to

Mortgage2.csv which were favourable. Like RF and the previous iterations of GBDT, they performed well on *Mortgage5.csv* which contains the synthetic samples. A GBDT with both 250 and 400 trees performed quite well, achieving accuracies of 64.9% and 65.7%, along with FPRs of 4% and 6% respectively. Once the number of trees was increased past 400, the FPR grew steadily as demonstrated in table 4.13.

	Mortgage4.csv			**Mortgage5.csv**		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	7.8	1	0	52.5	72.8	20
15	7.8	1	0	57	85.1	6
25	7.8	87	0	59.8	89.7	7
50	9.7	71.4	0	61.4	95.1	3
100	11.7	66.7	1	63.5	96.5	2
250	12.9	57.9	2	64.9	94.6	4
400	13.2	53.1	3	65.7	91.9	6
500	13.7	47.8	4	65.8	90.1	7
1000	16.4	44.7	5	71.5	83.9	14

Table 4.6 - GBDT results for clickstream only datasets

Adaboost performed extremely poorly on the *Mortgage4.csv* dataset, achieving a maximum accuracy of 13.7% with 1000 trees. This essentially means Adaboost is not a usable model for *Mortgage4.csv*. However the opposite is true for *Mortgage5.csv* where it obtained reasonable results. If the model was run with 50 tress or lower, it has a very high FPR, but once the number of trees was updated to be 100 or higher the FPR dropped significantly and accuracy increased slightly as can be seen in table 4.7.

	Mortgage4.csv			**Mortgage5.csv**		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	7.8	80	0	68.5	57.2	51
15	9.7	54.3	2	62.7	67.1	30.7
25	9.3	48	2	61.7	71.6	24
50	9.3	47.1	3	62.6	79.6	16
100	11.7	49.2	3	63.4	90.1	7
250	12.8	42.3	4	64.1	92.6	5
400	12.1	37.8	5	64	90.9	6

500	12.9	38.3	5	64	90.7	7
1000	13.7	38.9	5	64.2	89.9	7

Table 4.7- Adaboost results for clickstream only datasets

4.3 Results Evaluation

As mentioned in subsection *4.2.1*, the creation of a base model against which further models can be assessed is a common practice in predictive modelling projects. For this research, the decision tree which was created achieved a recall accuracy of 71.2% with a FPR of 6.4%. This is a relatively good model considering there is little tuning applied to it, only adjusting the minimum number of samples per leaf node. However the FPR was still relatively high with the decision trees incorrectly predicting approximately 230 accounts will enter arrears. Consequently accuracy alone should not be used as a measure of the predictive power of others models, both recall accuracy and FPR should be used in conjunction.

Following on from the establishment of the performance of the base decision tree, the next phase was to determine the performance of other modelling algorithms on the same data to determine which algorithms will be applied further in the research. Various different algorithms were analysed to assess their performance and it became clear ensemble models outperformed "traditional" modelling approaches such as LR and k-NN, as presented in subsection *4.2.1.1*. Ensemble models such as RF, GDBT and Adaboost performed well in the prediction task achieving accuracies of 71.8%, 76.5% and 74.3% and a FPR of 3.1%, 4.2% and 4.9% respectively. While RF achieved a marginally higher accuracy over the base decision tree, 71.8% to 71.2%, it had a substantially lower FPR of 3.1% compared to 6.4%. RF incorrectly predicted approximately 110 accounts will enter arrears, compared to 230 for the base decision tree. Of all the ensemble models, GBDT performed the best. While its FPR was higher than RF, its accuracy was far higher at 76.5% compared to 71.8%. Adaboost had a lower accuracy and a higher FPR than GBDT. However, it still outperformed the base decision trees and RF, which leads to it be the second highest performing model for this research. All three models provided additional predicting power over using a base decision tree, with GDBT providing a substantially higher performance of 76.5% compared to 71.2% accuracy while reducing the FPR by 2.2%.

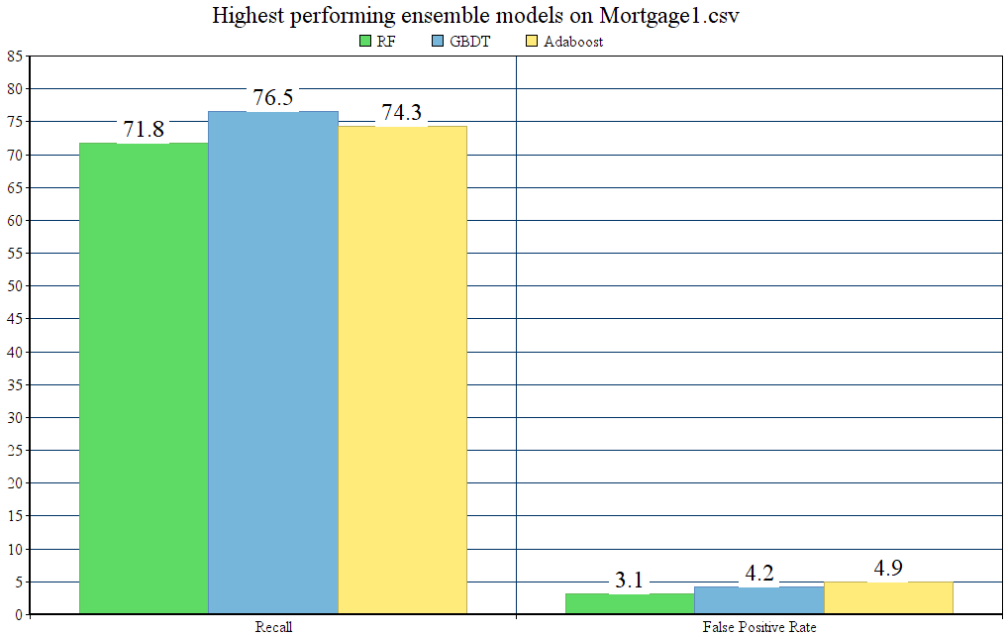

Figure 4.12 - Highest performing ensemble models for Mortgage1.csv

Therefore ensemble models were chosen as the models to be implemented in the next phases of the research which was assessing the impact the addition of clickstream features would have to the models performance. Initially, the three models were run against *Mortgage2.csv* and *Mortgage3.csv* with the clickstream features removed. This was to establish a further baseline against which the performance of models with clickstream features could be assessed. RF struggled to provide obtain any predictive power on *Mortgage2.csv*, achieving a maximum accuracy of 47.3% with 5 trees, which is lower than the no information rate and implies the model is worse than randomly choosing the arrears status of an account. This was caused by a very low sample sizes of *Mortgage2.csv* as when the same RF was applied to Mortgage3.csv it obtained as accuracy of 76.1% and a FPR of 1% which is a good result.

GBDT and Adaboost did not encounter the same issues as RF for *Mortgage2.csv* as they performed well, attaining a maximum accuracy of 67.2% and 65.2% and a FPR of 4% and 5% respectively. Like RF, GBDT also performed well on *Mortgage3.csv* hitting 74.1% accuracy and a FPR of 2%. Unfortunately the same is not true for Adaboost which had it lowest FPR at 6%, the same as the base decision tree while having an accuracy of 68% which is 3% lower than the base decision tree. Consequently GBDT is still the highest performing model.

Once the clickstream features were re-added back into *Mortgage2.csv* and *Mortgage3.csv*, the results were broadly similar to the datasets without the clickstream

features. Clickstream did push RF performance above the no information rate for *Mortgage2.csv*, albeit only slightly at 55.9% with a FPR of 3%. However this was still an increase of 11.8% which is significant. The opposite is true for *Mortgage3.csv* where the addition of clickstream features reduced the accuracy by 1.2% while the FPR remained the same. This is a marginal difference which may reduce (or increase) if the models were run again. Therefore it cannot be stated that clickstream features negatively or positively impact the performance of RF on these datasets as the results are inconclusive. The increase of 11.1% in *Mortgage2.csv* is encouraging, and may have proved to be even larger is there was a larger sample size available at the time of this research.

For *Mortgage2.csv*, the accuracy of GBDT increased from 67.2% to 70.3% with the addition of clickstream features, while the FPR increased by 1% to 6%. This relatively high FPR is detrimental to the performance of the model for the purposes of this research. Similar to RF, the accuracy of GBDT also decreased slightly by 0.8% for *Mortgage3.csv* with the addition of clickstream features, which is an insignificant amount. The FPR remained the same however. Adaboost performed similarly to GBDT for *Mortgage3.csv* where it's accuracy decreased by 0.4% from 72.8% to 72.4% while maintaining its FPR. For *Mortgage2.csv* however, Adaboost's accuracy remained the same, with its FPR increasing by 1% with the addition of clickstream features. Therefore as with before, it can be determined GBDT is the best algorithm for the purpose of this research.

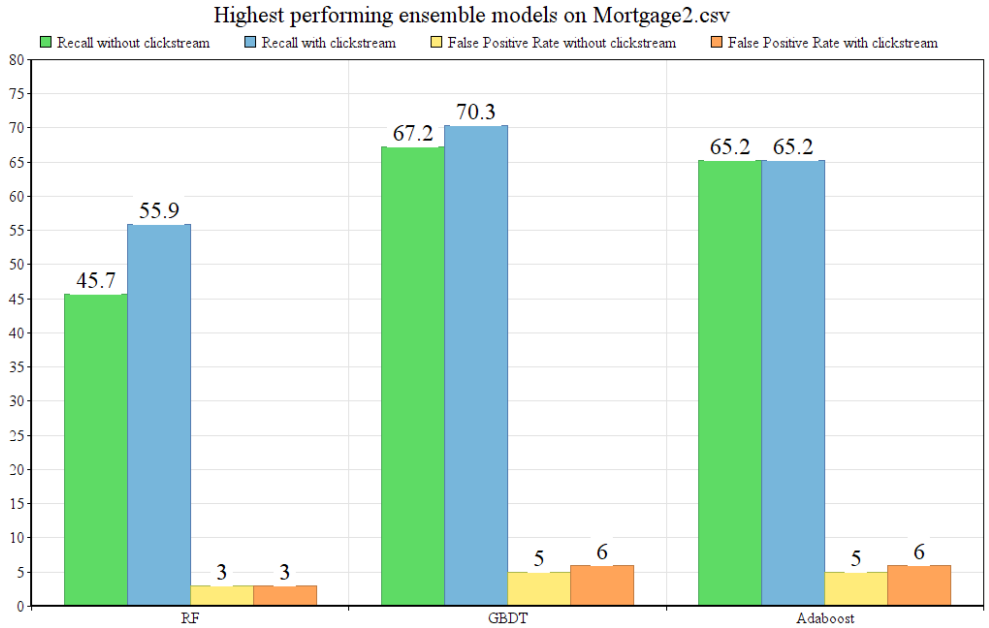

Figure 4.13 - Highest performing ensemble models for Mortgage2.csv

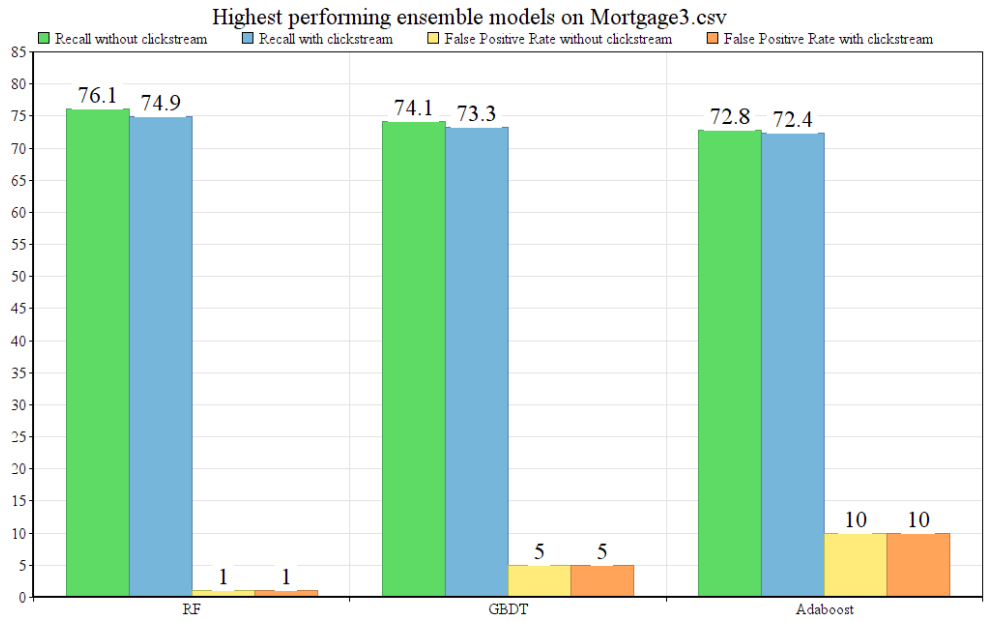

Figure 4.14 - Highest performing ensemble models for Mortgage3.csv

The final experiment undertaken was to establish the power of standalone clickstream features to predict the arrears status of a mortgage account. For this purpose, *Mortgage4.csv* and *Mortgage5.csv* were created in the same vein as *Mortgage2.csv* and *Mortgage3.csv*, in that the former contained the entire sample of accounts in arrears and the latter contained the same data in addition to extra accounts in arrears generated via SMOTE. Similar to the results from *Mortgage2.csv*, RF performed very poorly on *Mortgage4.csv* attaining a maximum accuracy of 17.2% with a FPR of 9%, which is an essentially useless model. This score was achieved with a RF of only 5 trees. When the tree count was increased to 250, as per the other high performing RF from previous experiments, the accuracy dropped to only 12.9%, as did the FPR rate which dropped to 4%, again a useless model. Nonetheless when the same models were run against *Mortgage5.csv*, the results seemed to be respectable. A RF with 250 trees obtained an accuracy of 72.1%. However, this model also as a substantially high FPR of 15% which is simply too high for this research.

GBDT performed very similar to RF on *Mortgage4.csv* only obtaining a maximum accuracy of 16.4% and a FPR of 4% with 1000 trees, which is a very poor model. For *Mortgage5.csv* however, they achieved a reasonable accuracy of 64.9% accuracy and a FPR of 4% with 250 trees. If the same 1000 tree model is applied, the accuracy increase quite a bit to 71.5%, however the FPR also increased significantly to 14%, again producing a useless model for this research. As somewhat expected, like the other ensemble models, Adaboost performed very poorly on *Mortgage4.csv* only attaining a

maximum accuracy of 13.7% and a FPR of 5% with 1000 trees. With the additional samples added synthetically to *Mortgage5.csv*, Adaboost achieved a respectable accuracy of 64.1% and a FPR of 5% with 250 trees, which is similar results to the rest of the ensemble models.

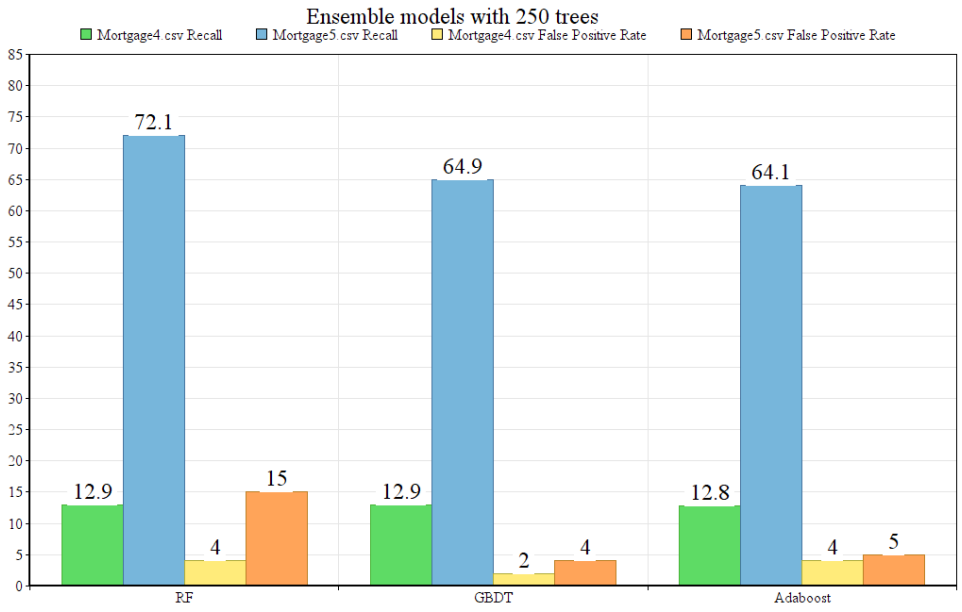

Figure 4.15 - Results of ensembles models with 250 on clickstream only features

Overall the models performed admirably once there was sufficient data for efficient and effective training. However, once the data didn't exist, the models struggled as is to be expected. GBDT performed consistently well across all datasets with the exception of *Mortgage4.csv*, due to the small sample size. The addition of the clickstream features for experimentation with *Mortgage2.csv* and *Mortgage3.csv* provided little impact in terms of accuracy for all models. For *Mortgage2.csv*, the addition of the features did improve accuracy across the board with the exception of Adaboost which remained static. Conversely, the addition of the features to *Mortgage3.csv* decreased the accuracy of all of the models, albeit only slightly with a difference of less than 1% for GBDT and Adaboost and a difference of 1.2% for RF. These numbers are close enough to not definitely state clickstream features had a detrimental impact on the performance of the models and they could potentially be different if the models were run again on the same data. A substantial issue with this experimentation is the size of the samples available for the research. All available data was utilized for this research.

4.4 Conclusion

This chapter presented the various experiments undertaken to determine the impact clickstream data can have on a mortgage arrears prediction model. To establish the power of traditional features, a base model was initially constructed. This base model was a decision tree which attained an accuracy of 71% with a FPR of 6%. From here, various other modelling algorithms were trained, tuned and tested on the various source datasets. Ensemble models proved to be the highest performing models with "traditional" models such as logistic regression and SVM performing poorly.

After all of the experiments, GBDT proved to be the highest performing models when recall accuracy and FPR are examined. For datasets without the synthetic samples added, random forests produced little predictive power while GBDT and Adaboost performed well. Like GBDT, Random Forests and Adaboost performed well on the datasets with synthetic records added. The final experiment involved establishing the predictive power of clickstream features. The same models were applied with traditional features removed, and produce somewhat decent results. The accuracy attained hovered around 65% across all models, which is more that the no information rate.

Finally all of the results are collated and presented and comparisons between the algorithms were presented and discussed.

5 CONCLUSION

5.1 Introduction

This chapter will conclude the dissertation and provide a summary of the key points describing how they relate to the research question and objectives set out at the start of this paper. Following this the contributions this research made to the body of knowledge will be discussed along with an evaluation of the experiments conducted. The results of the experiments are discussed along with limitations of the research. Potential areas for further research and additional experiments are then presented which is followed by concluding remarks.

5.2 Research Overview and Problem Definition

The research conducted for this dissertation involved reviewing modern analytical practices in the areas of two-class classification, clickstream data analysis and mortgage arrears with a vision of establishing the impact clickstream data can have on a mortgage arrears prediction model. Consequently this research drove the design and implementation of experiments to ascertain the predictive power of clickstream data for mortgage arrears prediction.

The below list is the objectives this research has achieved:

- Reviewed up-to-date literature on the Irish mortgage market focusing heavily on mortgage arrears in addition to reviewing modern approaches to data mining and skewed datasets.

- Designed a scalable approach to construct an ABT including the gathering and understanding of clickstream data and events.

- Implemented various predictive modelling algorithms and demonstrated the power of ensemble models for arrears prediction.

- Constructed a powerful base model which could be employed by Lender A's ASU team to determine which customer may be in financial difficult and require assistance.

- Established clickstream data can have a marginal positive impact on a mortgage arrears prediction model despite a very low sample size.

5.3 Contributions to the body of knowledge

The below points present the worthwhile contributions to the discipline:

- Demonstrated the creation of a model which could accurately predict mortgage arrears while also maintaining a low false positive rate. This model could be deployed by Lender A to assist the ASU team to help borrowers who may be in financial difficulty.

- Presented how clickstream data can be transformed to be utilised as input to a predictive model and also established the predictive power clickstream data can have with regards to mortgage arrears.

- Established various features which could be beneficial in the construction of a predictive model to accurately predict an arrears event occurring.

- Revealed the negative impact the inclusion of a customer's credit grade can have on a mortgage arrears prediction model.

- Observed the strength ensemble models have over traditional modelling techniques and also showed the predictive power of decision trees despite their simplicity.

5.4 Experimentation, Evaluation and Results

The primary goal of this dissertation was to ascertain if a customer's clickstream data can have an impact on the performance of a mortgage arrears prediction model. The initial step of the research was to transform the source data to allow for it to be employed as input features to a modelling algorithm. As the source of all of the data was Lender A's EDW, it was relatively clean and required little data preparation in terms of data cleansing. To allow for the data to be utilised in a predictive model, an ABT needed to be created which would contain one record per mortgage account. To create this ABT, all demographical, transactional and clickstream based features were aggregated to a singular record for the six month time span of the customer. This ABT contained 81,176 samples of which only 3,399 accounts were in arrears. Consequently the accounts in arrears only represented 4.2% of all of Lender A's mortgage accounts and thus the data

was heavily skewed. Therefore it was determined that some form of sampling would need to be employed to prevent the substantial skew towards non-arrears negatively impacting the performance of a model. To achieve this various datasets were created from the source ABT and these datasets are what was used an input to the modelling process. Each of these datasets contained a different split of source data in an attempt to capture as much information from source as possible. It was also at this stage where the very small number of mortgage accounts in arrears with valid clickstream data associated with them was established. There was only 851 accounts in arrears with valid clickstream data.

To establish if the addition of clickstream data could have an impact on the performance of a model, an arrears prediction model first needed to be constructed. Various different modelling algorithms were applied and ensemble models were the highest performing models, with GBDT attaining a recall of 76.5% on a base dataset without any clickstream features. Traditional predictive algorithms such as logistic regression and k-NN performed poorly as can be seen in subsection *4.2.1.1*. This original model is a good performing model and would be beneficial to the ASU team within Lender A to allow them to assist customers in financial difficulty. The next phase was to ascertain the predictive power of the clickstream features. To establish this, a second dataset was used due to the very small number of samples with valid clickstream data related to them. On this second dataset, a GBDT was again the highest performing model achieving similar results to the initial dataset with a recall of 67.2% without synthetic oversampling (lower than initial model due to substantially smaller sample size) and 74.1% with synthetic oversampling applied. With the addition of clickstream features the recall rose to 70.3% for the dataset without sampling applied. This is a reasonable increase and indicates clickstream data could prove to be beneficial to the model. However for the dataset with synthetic oversampling applied, the recall decreased slightly by 0.8% to 73.1% which in an insignificant amount and may not re-occur if the model was run again due to employing cross validation during the prediction process. The final experiment conducted was carried out to understand if clickstream data on their own could be used as an indicator for mortgage arrears. The results obtained demonstrated clickstream data was a moderately good source at determining arrears with models attaining a recall of approximately 65% while maintaining a lower false positive rate.

The results obtained by this research proved to be beneficial for Lender A. The model created for the base dataset could be deployed with little further work as the source data is already available in their EDW. However for now it would be advisable to not include clickstream data due to the very low number of accounts containing valid clickstream data. This issue could potentially be mitigated if the six month date range was increased as mentioned in the next section. Clickstream data did demonstrate its ability to increase a predictive models capability for determining arrears in certain circumstances. It is still very early days for clickstream data within Lender A's EDW which contains data dating back to 1998. As time moves on and more customers switch to online banking, clickstream data could become a pivotal source for not only predicting mortgage arrears but also arrears in various other loan sources such as personal and car loans. This research has demonstrated clickstream data, even with a very limited number of samples, can have a positive impact on a mortgage arrears prediction model albeit only marginally for now.

5.5 Future Work and Research

The sample of available clickstream data caused issues throughout this research which required oversampling to be employed in an attempt to alleviate this issue. Another potential solution for this issue would be to increase the date range of the available data. This research used a six month date range which in turn returns 851 accounts in arrears with clickstream data associated with them. If this period was increase in further research, it should produce more clickstream data. However the caveat with this is the availability of clickstream data only dates back to mid-2015 and consequently if a date range greater than 6 months was employed it would limit the starting dates from where the accounts could be taken (I.E. you cannot use accounts in arrears from early 2015 since the clickstream data simply won't exist). In addition to this, this research used Sept 2017 as the date to retrieve accounts not in arrears. If this date was moved to a later date this should increase the amount of accounts not in arrears which have valid clickstream data associated with them which could potentially reduce the number of false positives a model may produce.

As mentioned in subsection *1.5*, the classification of clickstream events is accomplished via pre-defined rules by Lender A. Some of these rules have not been updated since they were originally conceived mid-2015 when clickstream data capturing began. Consequently many of the events are being classified as the default "*Not Yet Classified*"

due to subtle changes in how both the clickstream data is being captured and the structure of the digital channels throughout the years. If these rules were updated it may better reflect the current state of Lender A's digital channels and therefore produce better results. In addition to this, further research could be conducted in an attempt to cluster these events to get an understanding of how they related to each other. This could produce event clusters which in-turn could be used in addition to singular events as input features to a modelling algorithm. This clustering was not carried out during this research due to time constraints. Another example of this potential clustering exercise is employing it to drive the targeting of relevant advertisements to customers.

For this research, only Lender A's mortgage accounts were examined. However Lender A also has subsidiaries who in-turn offer mortgages under their own brand name. The data associated with these mortgages are not currently in Lender A's EDW, but are planned to be integrated in the coming months. The addition of these mortgages could provide useful insights to the predictive modes created for this research. Also, the output of the mortgage default prediction models mentioned in subsection *1.1* may also prove to be a useful input feature to the arrears prediction models since a defaulted account had to be in arrears before defaulting.

Finally one further item which could be looked into is the analysis of BTL mortgages. While the number of BTL mortgages belonging to Lender A is substantially smaller than the number of PDH mortgages, there is potential for BTL mortgages to be examined. However careful consideration would be needed when determining what features to use as input as items such as the number of mortgages a person may have would need to be understood to prevent skewing of the data and results.

5.6 Conclusion

This chapter concludes the research undertaken to establish the impact clickstream data can have on a mortgage arrears prediction model. Within this chapter, a synopsis of the work carried out along with the key results of the experiments are presented. Potential areas for further research were discussed along with possible modifications to this research's experiments which may lead to an increase in the models predictive power. While clickstream data only had a small capability in increasing the predictive power of certain models these results are still positive considering the small sample size.

Consequently the outcome of these experiments can be deemed a success in that the addition of clickstream data has been shown to be potentially beneficial to the predictive capability of an arrears prediction model. The model constructed without clickstream features proved to be a versatile model with a reasonably high accuracy. Lender A's ASU team will be contacted and the results of this research will be presented to them to determine if and how they would like to progress

BIBLIOGRAPHY

Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random Forests and Decision Trees. *International Journal of Computer Science Issues*, 272-278. Retrieved March 20, 2018, from http://www.ijcsi.org/papers/IJCSI-9-5-3-272-278.pdf

Angelini, C. (2018). Regression Analysis. *Reference Module in Life Sciences*. doi:10.1016/B978-0-12-809633-8.20360-9

Banking & Payments Federation Ireland. (2015, September). Important information to help people in mortgage arrears. Dublin, Ireland: Banking & Payments Fereration Ireland. Retrieved April 5, 2018, from https://www.bankofireland.com/app/uploads/assets/online-mortgagesbpfiboi-mortgage-arrears-a5-lores.pdf

Barboza, F., Kimura, H., & Altman, E. (2017). Machine learning models and bankruptcy prediction. *Expert Systems With Applications*, 405-417.

Breiman, L., & Cutler, A. (2001). Random Forests. *Machine Learning*, 5–32. Retrieved March 14, 2018, from https://link.springer.com/article/10.1023/A:1010933404324

Bucklin, R., & Sismeiro, C. (2009). Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing. *Journal of Interactive Marketing*, 35-48. doi:10.1016/j.intmar.2008.10.004

Cai, T., & Guo, Z. (2016). *Accuracy assessment for high-dimensional linear regression.* Pennsylvania: University of Pennsylvania.

Camba, J., David, R., Betan, A., Lagman, A., & Caro, J. (2016). Student analytics using support vector machines. *Information, Intelligence, Systems & Applications (IISA), 2016 7th International Conference on.* Chalkidiki: IEEE. doi:10.1109/IISA.2016.7785425

Central Bank of Ireland. (2018, March 22). Residential Mortgage Arrears and Repossessions Statistics: Q4 2017. *Statistical Release*. Dublin, Ireland: Central Bank of Ireland. Retrieved April 4, 2018, from https://www.centralbank.ie/docs/default-source/statistics/data-and-analysis/credit-and-banking-statistics/mortgage-arrears/residential-mortgage-arrears-and-repossessions-statistics-december-2017.pdf

Chaudhuri, A., & De, K. (2011). Fuzzy Support Vector Machine for bankruptcy prediction. *Applied Soft Computing*, 2472-2486. doi:10.1016/j.asoc.2010.10.003

Cherkassky, V., & Ma, Y. (2004). Practical selection of SVM parameters and noise estimation for SVM regression. *Neural Networks*, 113-126.

Citizens Information. (2018, April 4). *Taking out a mortgage*. Retrieved from Citizens Information: http://www.citizensinformation.ie/en/housing/owning_a_home/help_with_buying_a_home/paying_for_a_home.html#l4e5b7

Cortes, C., Gonzalvo, X., Kuznetsov, V., Mohri, M., & Yang, S. (2017). *AdaNet: Adaptive Structural Learning of Artificial Neural Networks.* New York: AdaNet.

CSO. (2008). *Construction and House in Ireland 2008 Edition.* Dublin: CSO. Retrieved March 23, 2018, from http://www.cso.ie/en/media/csoie/releasespublications/documents/construction/current/constructhousing.pdf

Department of Housing, Planning and Local Government. (2018, June 1). *Department of Housing, Planning and Local Government*. Retrieved from Mortgage market statistics: http://www.housing.gov.ie/sites/default/files/attachments/4a-loan-app-num-all-by-year-1970-todate_2.xlsx

Dougherty, E., Hua, J., & Bittner, M. (2007). Validation of Computational Methods in Genomics. *Current Genomics*, 1-19.

EBS. (2017, January). Negative equity home movers: Guiding you through your next move. *EBS Mortgages*. Dublin, Ireland: EBS. Retrieved April 4, 2018, from https://www.ebs.ie/content/dam/ebs/pdfs/mortgage/ebs-negative-equity-home-movers.pdf

Flavin, T., & Connor, G. (2013). *Irish Mortgage Default Optionality.* Department of Economics, Finance & Accounting. Dublin: National Univeristy of Ireland, Maynooth. Retrieved March 22, 2018, from http://eprints.maynoothuniversity.ie/4584/1/N243-13.pdf

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). *Knowledge Discovery in Databases: An Overview.* Palo Alto, California: AI Magazine.

Gromski, P., Correa, E., Vaughan, A., Wedge, D., Turner, M., & Goodacre, R. (2009). A comparison of different chemometrics approaches for the robust classification of electronic nose data. *Analytical and Bioanalytical Chemistry*.

Hao, C., & Zhang, B. (2009). Review of the literature on credit risk modeling: Development of the recent 10 years .

Hara, A., & Hayashi, Y. (2015). Ensemble neural network rule extraction using Re-RX algorithm. *Neural Networks (IJCNN), The 2012 International Joint Conference on* (pp. 1-8). Brisbane: IEEE. doi:10.1109/IJCNN.2012.6252446

Hariri-Ardebili, M., & Pourkamali-Anaraki, F. (2018). Support vector machine based reliability analysis of concrete dams. *Soil Dynamics and Earthquake Engineering, 104*, 276-295. doi:10.1016/j.soildyn.2017.09.016

Ho, T. (1995). Random Decision Forests. *Proceedings of the Third International Conference on Document Analysis and Recognition*, 278-282. Retrieved March 15, 2018, from http://ieeexplore.ieee.org/document/598994

Hoens, T. R., & Chawla, N. V. (2013). Imbalanced Datasets: From Sampling to Classifiers. *Imbalanced Learning: Foundations, Algorithms, and Applications*, 43-59.

Huang, Z., & Luo, L. (2015, October 2). It takes the world to understand the brain. *Science Mag, 350, 6256*. AAAS.

Joachims, T., Finley, T., & John Yu, C.-N. (2009). Cutting-plane training of structural SVMs. *Machine Learning, 77*(1), 27-59.

Keeney, M., & O'Donnell, N. (2009). *Financial Capability:New Evidence for Ireland.* Dublin: Central Bank of Ireland.

Kelly, M. (2009). *The Irish Credit Bubble.* School of Economics. Dublin: UCD. Retrieved April 04, 2018, from https://www.ucd.ie/t4cms/wp09.32.pdf

Kelly, R., & O'Malley, T. (2016). The good, the bad and the impaired: A credit risk model of the Irish mortgage market. *Journal of Financial Stability*, 1-9. doi: 10.1016/j.jfs.2015.09.005

Kohavi , R., Yun , Y., & Friedman , J. H. (1996). Lazy Decision Trees. *AAAI-96 Proceedings*, 717-724. Retrieved March 15, 2018, from http://www.aaai.org/Papers/AAAI/1996/AAAI96-107.pdf

Kudlyak, M., & Ghent, A. C. (2011). *Recourse and Residential Mortgage Default: Evidence from U.S. States.* Review of Financial Studies. Richmond: Federal Reserve Bank of Richmond. Retrieved April 2, 2018, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=1432437

Lane, P. (2011, April 6th). The Euro Area Crisis and Ireland. Dublin. Retrieved April 6, 2018, from https://www.tcd.ie/policy-institute/assets/pdf/Lane_HenryGrattan_April11.pdf

Lane, P. (2017, December 11). Central Bank (Supervision and Enforcement) Act 2013 (Section 48) (Housing Loan Requirements) (Amendment) Regulations 2017. *Irish Statute Book*. Dublin, Ireland: Irish Government. Retrieved March 30, 2018, from http://www.irishstatutebook.ie/eli/2017/si/559/made/en/pdf

Lapedes, A., & Farber, R. (1988). *How Neural Nets Work.* Los Alamos: American Institute of Physics.

Leman, E. (2015, May). *Implementing the IFRS 9's Expected Loss Impairment Model: Challenges and Opportunities*. Retrieved from Moody's Analytics: https://www.moodysanalytics.com/risk-perspectives-magazine/risk-data-management/regulatory-spotlight/implementing-the-ifrs-9-expected-loss-impairment-model

Li, H., Wang, P., You, M., & Shen, C. (2018). Reading car license plates using deep neural networks. *Image and Vision Computing, 72*, 14-23. doi:10.1016/j.imavis.2018.02.002

Liu, H., & Yu, L. (2003). Feature Selection for High-Dimensional Data: A Fast Correlation-Based Filter Solution. In T. Fawcett, & N. Mishra (Ed.), *Twentieth International Conference on Machine Learning (ICML-2003)*, *2*, pp. 856-863. Washington DC. Retrieved April 8, 2018, from https://www.aaai.org/Papers/ICML/2003/ICML03-111.pdf

Longadge, R., Dongre, S. S., & Malik, L. (2013). Class Imbalance Problem in Data Mining: Review. *International Journal of Computer Science and Network*, 33-39.

McCarthy, Y. (2014). *Dis-entangling the mortgage arrears crisis: The role of the labour market, income volatility and housing equity.* Dublin: Central Bank of Ireland. Retrieved April 2, 2018, from https://centralbank.ie/docs/default-source/publications/research-technical-papers/research-technical-paper-02rt14.pdf

McCulloch, W., & Pitts, W. (1943). A logical calculus of the ideas immanent in nervous activity. *Bulletin of Mathematical Biophysics, 5*, 115-133.

Mehdikarimi, S., Norris, S., & Stalzer, C. (2015). *Regression Analysis of the Relationship between Income and Work Hours.* Georgia: Georgia Institute of Technology. Retrieved April 2, 2018, from https://smartech.gatech.edu/bitstream/handle/1853/53299/Regression_Analysis_of_the_Relationship_between_Income_and_Work_Hours-1.pdf

Norris, M., & Brooke, S. (2011). *Lift the Load: Help for people with mortgage arrears.* Department of Finance. Waterford: MABS. Retrieved March 28, 2018, from https://www.mabs.ie/downloads/reports_submissions/lifting_the_load_sep11.pdf

Pal, M. (2005). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, 217-222. doi:10.1080/01431160412331269698

Phangtriastu, M., Harefa, J., & Felita Tanoto, D. (2017). Comparison Between Neural Network and Support Vector Machine in Optical Character Recognition. *Procedia Computer Science*, 351-357.

Poel, D., & Buckinx, W. (2005). Predicting online-purchasing behaviour. *European Journal of Operational Research*, 557-575. doi:10.1016/j.ejor.2004.04.022

Quinlan, J. R. (1986). Induction of Decision Trees. *Machine Learning 1*, 81-106. Retrieved March 18, 2018, from http://hunch.net/~coms-4771/quinlan.pdf

Rakotomamonjy, A. (2004). *Optimizing Area Under Roc Curve with SVMs.* Saint-Étienne-du-Rouvray: INSA de Rouen.

Raphaeli, O., Goldstein, A., & Fink, L. (2017). Analyzing online consumer behavior in mobile and PC devices: A novel web usage mining approach. *Electronic Commerce Research and Applications*, 1-12. doi:10.1016/j.elerap.2017.09.003

Roche, J. (2014). Macroeconomic factors affecting mortgages. In *Predicting Mortgage Arrears: An Investigation Into the Predictive Capability of Customer Spending Patterns* (p. 16). Dublin: DIT.

Rodrígues, J. D., Pérez, A., & Lozano, J. A. (2010). Sensitivity Analysis of k-Fold Cross Validation in Prediction Error Estimation. *Transactions on Pattern Analysis and Machine Intelligence*, 569-575.

Rossum, G. (2009, Janurary 20). *A Brief Timeline of Python*. Retrieved from The History of Python: http://python-history.blogspot.ie/2009/01/brief-timeline-of-python.html

Sayad, D. (2018, March 18). *Saed Sayad.* Retrieved from About the Author: http://www.saedsayad.com/author.htm

Schoenmaker, D. (2015). Stabilising and Healing the Irish Banking System: Policy Lesson. *CBI-CEPR-IMF*. Dublin: Duisenberg School of Finance. Retrieved April 3, 2018, from https://www.imf.org/external/np/seminars/eng/2014/ireland/pdf/Schoenmaker_IrishBanking.pdf

Schorfheide, F., & Wolpin, K. (2011). *To Hold Out or Not to Hold Out.* Pennsylvania: University of Pennsylvania.

Shannon, C. E. (1948). A Mathematical Theory of Communication. *Bell System Technical Journal*, 379–423, 623–656. Retrieved from http://math.harvard.edu/~ctm/home/text/others/shannon/entropy/entropy.pdf

Sismeirob, C., & Bucklina, R. E. (2009). Click Here for Internet Insight: Advances in Clickstream Data Analysis in Marketing. *Journal of Interactive Marketing, 23*(1), 35-48.

Srivastava, J., Cooley, R., Deshpande, M., & Tan, P.-N. (2000). Web Usage Mining: Discovery and Applications of Usage Patterns from Web Data. *SIGKDD Explorations*, 12-23. doi:10.1145/846183.846188

Statista. (2018, March 22). *Online banking penetration in Great Britain from 2007 to 2017*. Retrieved from Statista: https://www.statista.com/statistics/286273/internet-banking-penetration-in-great-britain/

Taylor, C. (2018, Janurary 19). *The Irish Times*. Retrieved from Use of online banking more than doubled in Ireland in last 10 years: https://www.irishtimes.com/business/technology/use-of-online-banking-more-than-doubled-in-ireland-in-last-10-years-1.3360346

Tene, O. (2008). In *What Google Knows: Privacy and Internet Search Engines* (pp. 1434-1490). Tel Aviv: College of Management School of Law, Israel. Retrieved April 7, 2018, from http://boemund.dagstuhl.de/mat/Files/11/11061/11061.TeneOmer.Other.pdf

Vapnik, V., & Cortes, C. (1995). *Support-Vector Networks.* Boston: Kluwer Academic Publishers. Retrieved March 22, 2018, from http://image.diku.dk/imagecanon/material/cortes_vapnik95.pdf

Whelan, K. (2009). *Policy lessons from Ireland's latest depression.* School of Economics. Dublin: University College Dublin. Retrieved March 22, 2018, from http://www.ucd.ie/t4cms/wp09.14.pdf

Yang, Y., & Padmanabhan, B. (2003). Segmenting customer transactions using a pattern-based clustering approach. *AMCIS, III*(1), 411-418.

Yokota, T., Ishliyama, S., Teshima, S., Narushima, Y., Murata, K., Iwamoto, K., . . . Kikuchi, S. (2003). Lymph node metastasis as a significant prognostic factor in gastric cancer: a multiple logistic regression analysis. *Scandinavian Journal of Gastroenterology*, 380-384. doi:10.1080/00365520310008629

Zhang, C., Liu, C., Zhang, X., & Almpanidis, G. (2017). An up-to-date comparison of state-of-the-art classification algorithms. *Expert Systems with Applications, 82*, 128-150. doi:10.1016/j.eswa.2017.04.003

Zoppis, I., & Riccardo Dondi, G. (2018). *Kernel Methods: Support Vector Machines.* Milano: University of Milano-Bicocca. doi:10.1016/B978-0-12-809633-8.20342-7

APPENDIX A

	Mortgage2.csv			Mortgage3.csv		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	8	87.5	0	63.4	90.8	6
15	51.2	77.5	3	66.9	97.6	2
25	59	76.2	4	68.3	98	1
50	63.7	76.2	4	69.8	98.7	0
100	67.2	76.7	4	71.3	97.2	2
250	67.2	73.2	5	74.1	93.4	5
400	66	74.1	5	75.7	91.3	7
500	66.4	74.9	5	76.5	90	8
1000	64.8	75.4	5	77.3	89.2	9

Table A.1 - Results of GBDT on Mortgage2.csv and Mortgage3.csv with clickstream features removed

	Mortgage2.csv			Mortgage3.csv		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	1	1	0	61.2	90.5	6
15	57.8	78.3	3	64.2	97.1	2
25	66	77.9	4	65.8	98.4	1
50	69.5	76.4	4	67.6	98.9	0
100	69.1	74	5	70.1	97	2
250	70.3	73.2	6	73.3	94.2	5
400	71.1	72.8	6	76.1	91.9	7
500	71.1	74.3	5	76.3	91	8
1000	70.7	73.5	5	76.5	89.1	9

Table A.2 - Results of GBDT on Mortgage2.csv and Mortgage3.csv with clickstream features included

	Mortgage2.csv			Mortgage3.csv		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	54	70	5	62.6	70.9	25
15	59	70.2	5	66.1	87.5	9
25	59	71.6	5	68	91.8	6
50	65.2	74.2	5	70	90.4	7
100	65.2	73.6	5	72.8	87.6	10
250	64.5	71.7	5	74.1	85	13
400	62.5	71.5	5	74	84.6	14
500	63.7	71.9	5	74.3	83.7	15
1000	63.7	68.8	6	76	81.9	17

Table A.3 - Results of Adaboost on Mortgage2.csv and Mortgage3.csv with clickstream features removed

	Mortgage2.csv			Mortgage3.csv		
No of Trees	**Recall**	**Precision**	**FPR**	**Recall**	**Precision**	**FPR**
5	57.8	74	4	63.6	68.4	29
15	59.8	71.2	5	63.8	91.8	6
25	63.7	73.1	5	65.5	94.9	4
50	62.9	71.2	5	68.8	88.4	9
100	65.2	71.7	6	72.4	87.7	10
250	64.1	67.2	7	74.1	85.6	12
400	64.5	67.1	7	75.1	84.7	14
500	64.8	66.1	7	75.2	84.1	14
1000	65.2	65.8	7	77	81.8	17

Table A.4 - Results of Adaboost on Mortgage2.csv and Mortgage3.csv with clickstream features included