Enhancing Credit Analysis and Assessment using Geo Spatial Techniques

Deepak Kumar Gupta, B.Tech. Computer Science
Shruti Goyal, B.Tech. Instrumentation and Control

A Practicum submitted to University College Dublin in part fulfilment of the requirements of the degree of M.Sc. in Business Analytics

Michael Smurfit Graduate School of Business, University College Dublin

September, 2017

Supervisors: Dr. Peter Keenan, UCD
Selwyn Hearns, KPMG IRM Audit

Head of School: Professor Ciarán Ó hÓgartaigh

Dedication

To our freinds and family for their support and encouragement.

Contents

List of Figures

List of Tables

Preface

> *Live as if you were to die tomorrow. Learn as if you were to live forever.*
>
> — Mahatma Gandhi

The basis for this practicum stemmed from our passion towards data exploration, pattern finding, and development of new solutions. Data visualization has always been of our interest. Hence, this practicum presents an interactive visualization. It is our passion to develop interactive dashboards and serves profound insights to all types of users. We could not have achieved the success of our practicum without the constant help and support from our academic supervisor Dr. Peter Keenan and business supervisor Mr. Selwyn Hearns. We faced challenges while doing the research, but doing aggressive investigation has helped us to find answers to our questions. Research of this practicum has given us an opportunity to broaden our knowledge area towards data analysis and optimization of problems.

Chapter 1 will provide an overview of the problem this practicum is going to address with brief outline of each chapter.

Chapter 2 will discuss the business we are working with and what are the main contributions.

In chapter 3 detailed review of academic contributions towards the problem areas can be found. It will explain technologies that have been used in the past and are currently in use. Comparison between traditional systems and advanced methods can be of interest for the readers.

Chapter 4 will describe detailed steps of the methods followed while working on this practicum such as implementation, data discovery, etc.

Chapter 5 will presents results of the experiments that have been performed during the whole course of the completion of this practicum.

Data analysis and patterns that we have found are explained in chapter 6 and chapter 7 will provide closing statements and the scope of improvements in future.

The whole process could not have been achieved without a supporting group. At first, our family, who have always encouraged us with love and second, our supervisors, who have been so patient and provided their guidance throughout the practicum process. Thank you our support system for constant support.

University College Dublin — Deepak Kumar Gupta
August 31, 2017 — Shruti Goyal

Acknowledgements

We would like to express our thankfulness to Dr Peter Keenan and Dr James McDermott for constant support and providing answers all our queries related to this work.

We also like to thank KPMG, Ireland for sponsoring this project. Mr Selwyn Hearnes, Partner and Mr James Fitzpatrick from KPMG IRM Audit team for their valuable knowledge and support for carrying out this research work.

Abstract

This practicum presents interactive dashboard for enhancing credit analysis and assessment using GeoSpatial techniques for Irish residential properties developed by data from following sources (i) loan portfolios and credit data (ii) property price register (iii) Central Statistical Office. These analyses can be used to reduce the chances of financial loss on a residential mortgage.

The predictive model is built on logistic regression and decision tree algorithms and produces an estimated default probability of the applicant. Models are built on normalised data to cover all possible scenarios from real life. The predictive probability will determine good and bad customers by classifying them into four categories. The output from predictive models is used on tableau to generate business dashboard. Models ability to classify and performance measurements were measured by using statistical metrics: Gini, KS and AUROC. Results showed that decision tree has much better performance than logistic regression.

Keywords: Credit Scoring, Logistic Regression, Decision Tree, Tableau, Auditing

Executive Summary

Irish property crash (2007-2010) affected the financial stability of banks, and since then it is required to provide an intense focus on loans performance especially for residential loans to prevent losses in future. Also, there has been a significant improvement and enhancement in the technology used that allow interactive and detailed credit assessment. Financial data such as average property sales price, unemployment rates, average salary, interest rates are readily available along with Geospatial data such as property latitude and longitude. This practicum is focussed on assisting auditors and analysts by providing greater coverage of loan portfolios and problem areas by using Geospatial techniques, predictive modelling (Logistic Regression and Decision Trees) and visualisation tools such as Tableau.

Table 1: Test results for Logistic Regression and Decision Tree performance

	Logistic Regression			**Decision Tree**		
Data/Measure	**AUROC**	**KS**	**Gini**	**AUROC**	**KS**	**Gini**
Normalized Data	67.61	24	16	81.4	59.96	62.8

Data set is artificially generated by the business to comply with the confidentiality agreement with the clients. Provided data consists information of 237K loan accounts and thirty-five characteristic variables such as *CreditRating*, *LoanToValueRatio*, *LoanBalance*, *ProbationaryLoans*. To maintain scalability of data for future enhancement data is processed on Alteryx. To predict the probability of default, two models are trained and tested on logistic regression and decision tree(CART) algorithm. Performance of both the models is tested on AUROC, KS and Gini metrics as shown in Table.1. It can be seen that Decision tree outperforms Logistic regression and gives better accuracy.

The dashboard allows the user to choose an origin city and desired distance of the nearby area to analyse a selective number of accounts with predicted default probability shown in fig. 0.1.

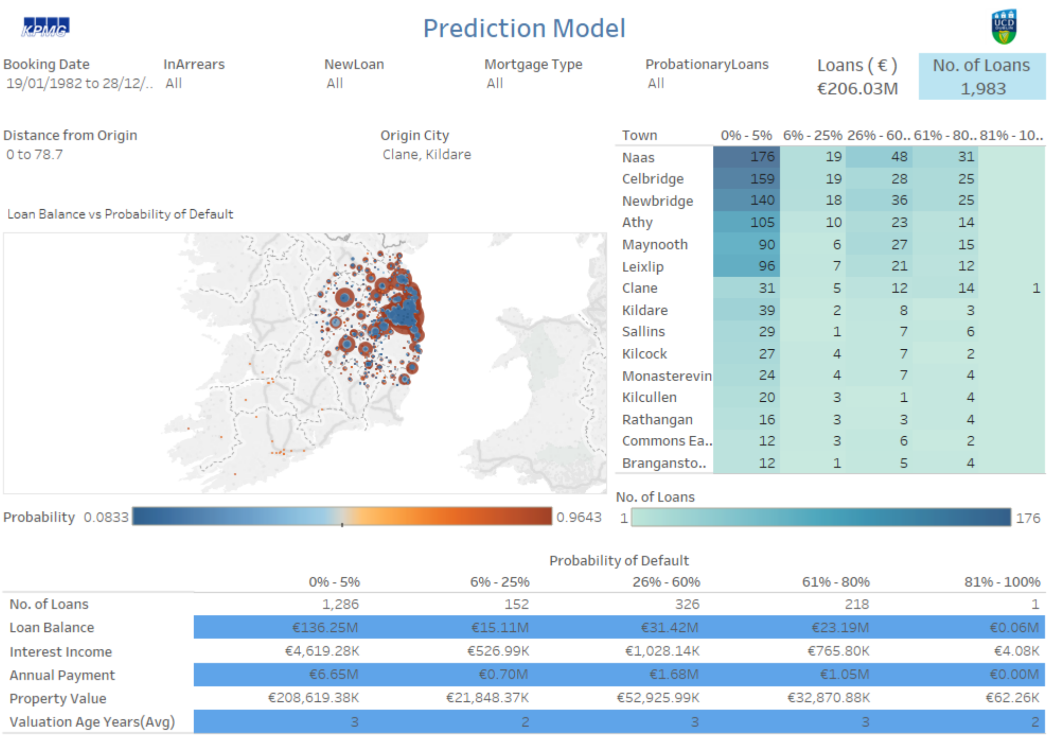


Figure 0.1: Predictive Dashboard built on Tableau

Chapter 1

Introduction

One of the key activities of banking and financial institutions that enhance their quality and financial system, correct handling, and management of liabilities. The performance of those tasks is very crucial for country's economic development, that Irish government witnessed as Irish property bubble that happened in Celtic Tiger period (late 1990 - 2007). While assessing credit risk, it is essential to validate the accuracy and reliability of credit scores or credit rating for all participants. So, How do banks identify a default event: 1. Non-repayment of the debt to the bank, 2. Repayment is due for more than 90days.

This work will discuss predictive models for enhancement in credit analysis and assessment of residential mortgages registered in Ireland using geospatial locations. There are many studies and researches on how to assess and analyse credit scoring or credit risk, but very few studies are present that describes assessment using geospatial data. This project will demonstrate how geospatial techniques can be used to enhance further credit analyses that empowers banks and financial institutions to take the much better decision on an application. This project will present a predictive model that predicts the probability of default and an interactive visualisation highly focused on geospatial locations of residences registered in Ireland and bank's branch locations. The purpose of this visualisation is to support decision maker to take a more efficient decision whether to provide loan on a particular house mortgage or not with the use of predicted probability of default. Models for Credit analysis are developed with the use of decision trees using CART algorithm and logistic regression for binary response (dependent) variables. While building models, potential variables were selected based on Information Value statistics. Credibility and quality of the models were evaluated using approaches such as GINI statistics, prediction accuracy, and ROC (Receiver Operating Characteristic) curve.

Credit assessment and analysis plays a crucial role in determining the financial strength of businesses and risk estimation that are associated with credit. Following are the primary purposes of assessment of credit:

1. Helps to keep track of the economy (macro economic perception)
2. Analyses and ensures stability of financial market (macro prudential perspective)
3. Assessment of quality of collateral/mortgage (monetary policy)

1.1 Assumptions & Challenges

KPMG provided made up data due to a confidentiality agreement with their client. Data is generated from pre defined formulas that made data look like original real life data, but it could not cover all possible real life scenarios. For example - Data only considers that an applicant will default if it has a credit rating of 5 but data did not consider the situation that a claimant may default if heshe has a credit score of 2,3,4 and even 1 in some cases. This case depicts a constraint of given data over real life data.

Below is a list of assumptions undertaken during the process of practicum:

1. Property prices have been considered as provided in the data by KPMG; there is no consideration of any time frame. For example, the date when property valuation was done.

2. Geospatial data such as address latitude and address longitude is assumed to depict geospatial location property correctly.

3. A property is considered as a whole, some apartments and number of floors are ignored. What latitude and longitude of a house consist of 2 floors are same.

4. Dimensions of house and size of the house(number of rooms, bathrooms, lawn, etc.) are not considered during model development.

5. This project only focuses on residential properties, not on commercial properties.

6. This project did not consider factors such as neighbourhood, amenities, and demographics which affects the property price in the market. However, factors such as location, average price have been considered for predicting the probability of default.

1.2 Outline

Below is the flow of the practicum which will give a brief description of each chapter:

- Business Background
  This chapter describes business need and contributions in detail. It will explain how this project will contribute towards banks and financial institutions businesses.

- Literature Review
  Chapter 3 presents an in-depth study of academic contributions achieved in the field of credit analysis, geospatial techniques, and data visualisation. This section will explain in detail what is credit scoring and what methods have been used in the past to enhance assessment of credit. It will show a comparison between traditional systems and credit scoring along with algorithms to build a model for predicting the probability of

default. Later, it will describe geospatial techniques and data visualisation techniques.

- Methodology
  Chapter 4 will give a detailed explanation of steps and tools that have been used to successfully conduct this project and how different tools have been integrated together.

- Results
  Chapter 5 explains the output generated from the methods and algorithms described in the sections mentioned above. It will describe the graphs and images that hold uttermost importance and are relevant to the business need along with Tableau dashboards.

- Discussion
  This chapter will discuss data limitations and practicality of the models developed that correctly answers business questions.

- Conclusion and Future Work
  This chapter will conclude the outcome of the practicum along with the improvements and future scope of the project.

Chapter 2

Business Background

2.1 Introduction

KPMG is one of the most renowned Big Four auditors and provides tax, audit, advisory and consultancy services to various clients. Information Risk Management is the service line of the organisation that provides information systems security assurance while minimising risks and frauds. For accuracy of financial reports, IT organisations depend on an effective audit. KPMG's IRM audit team works with clients and auditors to assist them to obtain their desired results; by assuring customers how IT functions are efficiently controlled and by ensuring auditors that their work is efficient and accurate within the guidelines. IRM audit team supports audit planning process and fraud risk assessment to monitor IT risks; supervises processes for a particular industry; supports auditors; assesses application controls design; supports testing phase of the whole audit process. Benefits of the services provided by IRM audit team are efficient and effective audits, impactful audit decisions and opinions, precise identification of business risks and issues reporting to senior management and audit committee.

2.2 Business Contribution

There has been a rapid loan growth since last few decades, which led to aggressive lending(weak controls and lenient standards). This increased lending can come from a volatile source. Auditing loan portfolios are imperative to make sure safety and compliance with regulatory requirements.The objective of auditing is to find errors and issues and take appropriate corrective measures or actions. Auditing of residential loan portfolios can alert users and banks about the deviations in prescribed policies of credit risks and therefore maintains sustainability and profits of banks. As mentioned in chapter 1 since the Irish property bubble in 2007-2010, the focus has been increased on the performance of loan portfolios especially in residential sector to achieve:

- Interactive way to identify patterns in datasets to drill down into problem areas
- Well timed potential issues indicators that adhere to provisions of audit processes and assessment of residential loans
- Better and greater coverage of problem areas and increased focus on judgemental loan applications
- Integration of useful and relevant market data and economic indicators for enhanced loan assessment

There has been a significant improvement in technology that helps in analysing data interactively and graphically. Growth in financial services has led to an increase in accuracy of loan data and better availability of external data sources. This practicum will bring together such information in an interactive way to enhance credit analysis, audit and assessment of residential loan portfolios to reduce the cost of credit analysis, enable faster credit decisions, close monitoring of accounts and prioritise collections.

Chapter 3

Literature Review

3.1 Introduction

In recent years, the purchasing power of an individual has increased due to an economic boom which further resulted in more employment, better wages and decline in inflation rates. All these factors empower a consumer to purchase new commodities for short term as well for long term investment goals. In long term, consumer tend to invest in real estate and to achieve that purpose customer approaches financial institutions or banks to seek monetary help in term of credit or loans.

Suppose, a client wants to buy a new car, but he/she does not have access to sufficient funds to make full payment. Also, he/she will not be able to pay the full or partial amount through his/her credit card. These circumstances can occur any time, where one may need a certain sum of money. So one needs to borrow a generous amount of money from some other entity which is called a loan. A loan is lending a sum of money from one entity to another that involves repayment of the amount in the near future. The Lent amount is called principal amount and amount to be repaid a summation of principal amount and an interest amount or other charges. It is not as easy as it sounds like, there are certain terms need to be agreed upon by each entity before exchange of the money. A loan can be for an amount taken at one time or can be made in instalments Partial Payments]. A loan can be provided by banks, corporations and financial institutions. Banks and financial institutions offer various types of loans as per the need of an applicant, such as personal loans, home loans, business loans, credit card loans and cash advances. There are times when the borrowing amount is enormous, and banks cannot provide the credit based on verbal agreement, and they need to ensure that if an applicant is not able to repay the loan, then they need to have a source to recover the lent amount. So, in this case, an applicant needs to apply for a mortgage with the bank.

A mortgage or collateral is an instrument that applicant has to pay back with predefined series of payments to the bank and financial institutions. Over a duration of time, an applicant needs to repay the loan inclusive of interest amount to free his/her mortgage. In case, if an applicant is not able to repay the loan within a predetermined time, then the bank can recover their money by selling or putting it for auction the mortgage. The most common type of mortgage is residential mortgages were applicant gives his/her house to banks, and in a case of no repayment then a bank will claim the house to recover the balance amount of the loan. This will give a bank a security that their lent amount is not at risk and over the years they will get back their lent money one way or the other. Mortgages come in various forms. Most commonly used mortgage types are Fixed Rate Mortgage where applicant repays the loan amount on a fixed rate throughout the period determined and Adjustable Rate Mortgage where interest rate varies as per the changes in market interest rates. Our work is based on analysis of residential mortgages with varied interest types which will be discussed in later sections.

Before analysing data based on residential mortgages, one needs to understand the process of giving a loan. Depending upon the requirement an applicant applies for a loan by filling an application form with all the necessary details required by the bank. Bank officials then analyse the application and may

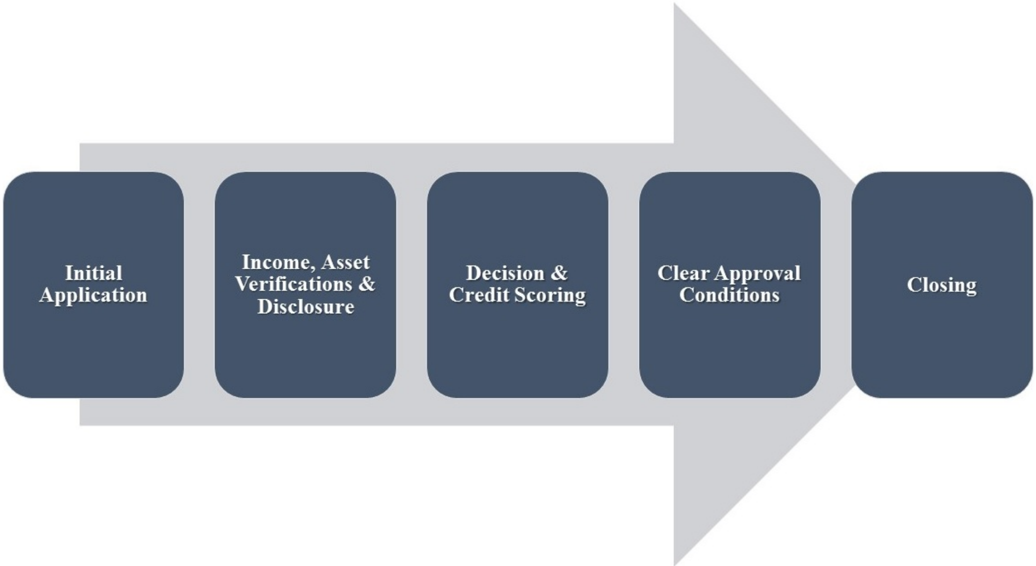


Figure 3.1: Loan application flow chart

Source: Designed using MS Office template

ask an applicant for additional information; after evaluation, bank approves or disapproves the loan. Next, borrower and bank sign an agreement that states all the terms and conditions of the loan including determined interest rate and type of mortgage. Lastly, loan amount will disburse, and borrower will start repaying the instalments that constitute principal amount and interest amount for predetermined period.

And, the major question is how do banks decide whether to give a loan or not? This issue is of primary concern as bank's cash flow highly depends on timely repayment of the loan. Every bank does not have the same procedure, but majority of the loan review process is same. Following are few characteristics that bank officials will concentrate while evaluating a loan application:

1. Credit history of applicant
2. Loan to Value ratio
3. Employment history

4. Character assessment of applicant

5. Evaluation of collateral

6. Financial statements such as bank history, cash flow, etc.

3.2 What is Credit Scoring?

One of the most important questions of borrowing and lending process of loan is How do banks make sure whether to give a loan to a borrower or not? Banks do credit evaluation of an application to make credit management decisions. Officials collect, analyse and classify credit variables and elements to reach credit decisions. Credit evaluation determines the quality of the bank. A process of evaluating customer's bad credit risk is called credit scoring. Since ages, there have been various definitions of credit scoring; Hand and Jacka (1998) stated that credit scoring is a process of measuring customer's creditworthiness. Anderson (2007) segregated credit scoring into two components: credit that means you can purchase now and repay the amount later; and, scoring means ranking based on predefined set of qualities to differentiate amongst cases to achieve credit decisions. On the other hand, Gup and Kolari (2005) stated that process of credit scoring uses statistical approaches to determine whether a borrower will default in future or not. Similarly, Beynon (2005) said, credit scoring is a statistical model that convert relevant credit data into numerical data that support credit decisions.Credit scoring techniques have been widely used to access commercial loans, businesses, real estate industry and residential mortgages (Gup and Kolari, 2005). Credit scoring is a method that decides whether an applicant will get credit, what will the process of getting credit and how will the strategies enhance borrower's profitability. Credit scoring models are prevalent from last ten decades that has evaluated consumer credit secure and reliable (Thomas *et al.*, 2002).

3.2.1 Traditional Subjective Assessment System and Credit Scoring

The primary objective of credit evaluation process is to compare and contrast characteristics of an applicant with other previous candidates who have repaid the loan amount. Bank will check candidate's profile with earlier candidates, if a profile is very much similar, then they will check if an applicant has repaid the loan on time. If a claimant did not default then the loan can be granted, if not then loan application will be rejected. Crook (1996) stated that there are two techniques for credit evaluation: Credit Scoring and Officials Subjective Assessment. Traditional judgement assessment method is entirely dependent on evaluator's experience and knowledge (Sullivan, 1981; Bailey, 2004). Subjective assessment is subjective and inconsistent, but on the other hand it can be successful, creditor's experience can be qualitative that helps in taking successful credit decisions.

While in credit scoring method, creditors use their knowledge and historical information of the loan applications to form an evaluation model to determine creditworthiness. Credit scoring methods are consistent, and self-operated that includes quantitative measurements of applicant's credit score subjected to predictor variables such as employment duration or credit history. Also, credit scoring method provides an advantage to a bank to keep their good credit customers intact and to improve customer service. Consequently, this process has been criticised because data that has been used consists of some assumptions to evolve model statistically.

3.2.2 Advantages and Disadvantages of Credit Scoring

Crook (1996) said that credit scoring process does not require too much information because the process the model has been statistically developed for a particular set of variables; on the other hand, subjective assessment does not have any variable reduction method because of no statistical importance.

Credit scoring method reduces bias by inspecting rejected applications; it will keep score how rejected applicants would have behaved if they have given the loan. It considered both good and bad credit players and built a model on a large number of applications compared to traditional methods. Scoring models also contain a significant number of relevant variables that show a correlation between variables and payment behaviour. A great advantage of this approach is its re-usability; the process can be used multiple time over the same data set with accuracy. Scoring models reduce processing cost and time with efficiency and ease decision-making process.

But, at times credit scoring model can inaccurately predict the creditworthiness of an applicant because of misclassification error. Due to its variable reduction technique, a model can miss out important variables to evaluate application which can be necessary. There may be chances that an applicant can repay the loan on time but based on the historical data or any missing information; a model can predict the wrong result. Also, these model can not be standardised as each industry can have different credit scoring models. Historical data can play a disadvantage as due to advancements in technology and rapid changes in economic factors, credit score model prediction can be inaccurate. Models are standardised and need to update as per the economic factors, that can cost much, and the process is not easy.

3.2.3 Is credit scoring process optimal?

Despite so much criticism on credit scoring models performance, credit scoring models are in use. But, there are some open questions which have left unanswered: Optimal evaluation of an applicant, relevant variables to evolve a model, information needed to enhance decision making, best measures that can predict loan accuracy, extent to which an applicant can be classified as defaulter (Al Amari, 2002).

Contrast to Al Amari (2002) questions, Abdou (2009) added more open questions to credit scoring process: How to choose appropriate technique to perform classification? Are there any other better classification methods better than credit scoring method? Is predicted value of the credit scoring model efficient than other methods? How to find out appropriate factors that influence credit scoring?

As mentioned above that credit risk majorly enhance bank's quality in spite of economic and environmental changes. So banks need to have suitable methods to evaluate credit risk. A good system should be able to correctly classify between good and bad credit customers because bad credit could cause some severe issues to the bank. Our work will discuss few techniques that can be used to evaluate credit risk by determining a probability of default and classification of chances of default. Also, our work will try to find out techniques that can enhance the assessment and analysis process of the credit.

3.3 Analysis and assessment of credit

Importance of assessing credit worthiness has been increased since, the property crash in 2008. Banks and Financial institutions making efforts to enhance traditional credit scoring mechanisms by incorporating latest technology and tools. Not only availability data about the customer but also rapid development in machine learning and analytics providing a foundation stone to banks.

Traditional credit scoring process with the random selection of good and bad portfolio from creditors file around 50 - 300 Capon (1982) characteristics points from loan portfolios to build an essential subset to perform statistical analysis. Hand and Henley (1997) mentioned about three commonly used approaches used for selecting characteristics out of available data: Expert Knowledge, Stepwise Statistical procedure and evaluating individual characteristics. Subject Matter Expert(SME)

Credit analysis and assessment are very important for banks and financial institutions to evaluate the credit worthiness of an applicant or a borrower. Banks implement various factors while assessing credit risk; such as credit rating, loan to value ratio, the probability of default, etc.; that leads to the derivation of credit risk rating. Variety of financial techniques have been used by the officials to analyse credit risk.

An applicant credit score is generated using credit rating system based on various characteristics points. After that credit score is used depending on the usage of the system. There are single cut-off and two cut-off stages in deciding application decision. In single cut-off, credit is granted if applicant score is higher than cut-off; otherwise, credit is denied. Some institutions incorporate two stage cut-offs, in this system if credit score is greater than upper cut-off then credit is granted and denied if the score is lower than lower cut-off. If the score is between upper and lower cut-off, then applicant credit history is pulled to calculate further scoring point and added to credit score. If new total score is greater than upper cut-off, then credit is granted else denied.

Banks and financial institution sets their cut-off for credit score based on the probabilities of each applicant ability to repay or non-payment of credit amount. Credit Risk has received a lot of criticism as well from Academics and Researchers. Al Amari (2002) has questioned about optimal method to evaluate customers? What are key variables or data points which an analyst must consider while evaluating customer applications? On what basis one can classify an applicant as good or bad?

However, apart from above questions following can be useful when building a new credit scoring system. One should evaluate statistical techniques or algorithm by its accuracy to correctly classify historical portfolios into good or bad credit from creditors file. Also, Banks and Financial institution's identified factors that can influence the prediction of credit and loan quality by gathering all possible information from customer applications form, bank transactions history and previous credit history. Credit Analysts analysis of all these information to decide what all variables or characteristics to be included in final the credit model.

One of the principal objectives of credit scoring system is to assist Banks and Financial Institutions to streamline their credit management procedure and policy that will enable analysts with an efficient tool which will provide fast and accurate analysis of credit.On the longer run, such tool helps banks to avoid bad credit and scale up bank revenues and profit by selling more financial products to customers.

3.4 Diffrrerent Technology in Credit Risk:

Table 3.1: Different Statistical Algorithms for Credit Scoring

Method	**Authors**
Linear Regression	Lee and Chen (2005); Hand and Henley (1997)
Discriminant Analysis	Fisher (1936); Durand *et al.* (1941); Altman (1968); Eisenbeis (1978); Zhou *et al.* (2016); Liberati *et al.* (2017)
Logistic Regression	Hosmer *et al.* (1989); Altland (1999); Nie *et al.* (2011); Abdou *et al.* (2008); Bensic *et al.* (2005); Joanes (1993)
Decision trees	Kohavi and Quinlan (2002); Breiman *et al.* (1984); Zhang *et al.* (2010); Zekic-Susac *et al.* (2004); Zhou *et al.* (2008); Huang *et al.* (2007); Xia *et al.* (2017); Koh *et al.* (2015); Koutanaei *et al.* (2015)
Neural networks	Demuth *et al.* (2008); West (2000); Gately (1995); Presky *et al.* (1996); Ghosh and Reilly (1994); Desai *et al.* (1996)

Linear Regression allows one to build to simple model using a dependent and two or more predictor data points, and it is being used in credit scoring models as the two class problems can be represented using a dummy variable (Lee and Chen, 2005). A Poisson regression can be used to classify cases where customer tends to partial repayments, and these payments can represent as a Poisson count in the model. Credit analysts can promptly analyse using linear regression credit model to investigate customer factor such as past payments record, credit guarantees and default, etc. against a predefined cut-off credit score. If new applicant credit score is higher than cut-off score, then credit is granted (Hand and Henley, 1997).

Discriminant Analysis: In credit scoring models, a statistical analysis method called Discriminant Analysis is regularly used by the researcher to rapidly build a prototype model when there are two or more categorical dependent variables for analysis. Multiple Discriminant Analysis(MDA) utilised in various studies and business verticles for the variety of applications since its inception in 1930's (Fisher, 1936). Durand *et al.* (1941) used the Discriminant analysis for modelling a scoring system that gives a prediction about loan repayment. Many researchers agreed that the MDA is the best use to classify a group of categorical variables into two or more predictor or classes. For example, Credit Analyst can build a scoring system using MDA to categorised a new loan application into Default or Non-Default category, and this will help banks to avoid those applicants who have potential to default in repayment sooner or later. Altman (1968) used MDA by developing a scoring model based on five financial ratios by analysing financial statements to select eight variables for predicting financial bankruptcy in Corporates. Eisenbeis (1978) noted the problem associated with Discriminant Analysis such as reduction in dimensionality, improper estimation of classification error, using linear functions instead of quadratic functions, etc. Despite these limitations in MDA, it is still one of the techniques which are often used by credit analyst in building credit scoring system (Zhou *et al.*, 2016; Liberati *et al.*, 2017).

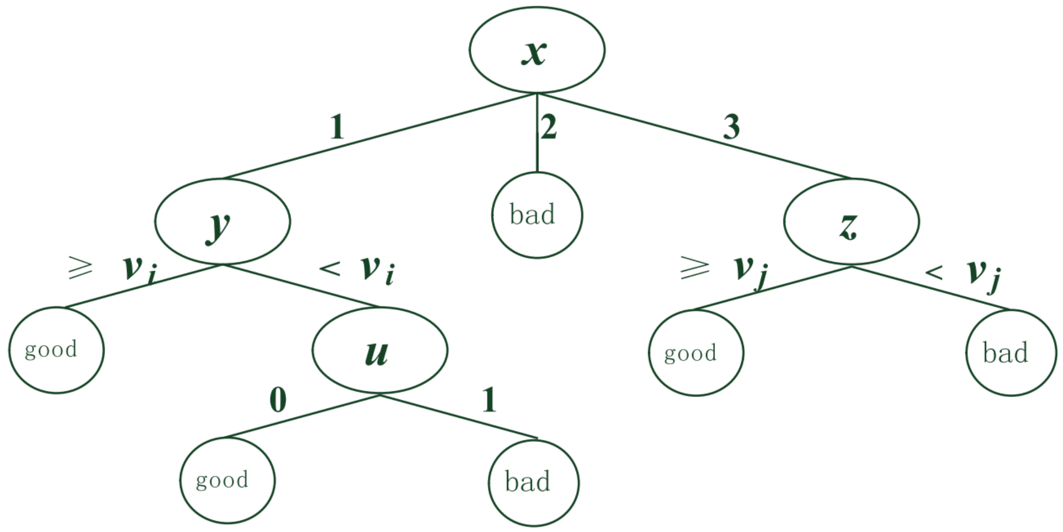


Figure 3.2: Simple Decision Tree Source: (Zhang *et al.*, 2010)

Logistic Regression has resemblance with Linear regression and it is also most commonly used statistical technique for building credit scoring system. Dichotomous nature of logistic regression outcome probability (good credit or bad credit) makes it different from linear regression (Hosmer *et al.*, 1989). By using two or more independent variables, one can build the simple logistic regression model. However, logistic regressions with more than one independent variables use the maximum likelihood method to build credit scoring model (Altland, 1999). Logistic regression has been widely used in building credit scoring system in financial domain (see for example: Nie *et al.* (2011); Abdou *et al.* (2008); Bensic *et al.* (2005); Joanes (1993))

Decision trees is one of the classification technique in machine learning and widely using for building credit scoring system. Classification & Regression Trees (CART) and C4.5 are two widely used decision tree algorithms (Kohavi and Quinlan, 2002). One of the first model pioneered by Breiman *et al.* (1984). With the help of single input function, algorithm splits all data observations to generate a dichotomous tree using CART. The algorithm chooses the best subset data based on the lowest cost of misclassification (Zekic-Susac *et al.*, 2004). This process of selecting an attribute from data subset is repeated as algorithm C4.5 or CART continues to choose one attribute that splits data into subset based on information gain (Zhou *et al.*, 2008). Huang *et al.* (2007) used decision tree along with support vector machines to build credit scoring model. Other applications on using decision tree in credit scoring has been discussed by (Xia *et al.*, 2017; Koh *et al.*, 2015; Koutanaei *et al.*, 2015).

Neural networks in machine learning or data mining is modelling system, which is based on the human brain and nervous system. A Neural network consists of several neurons(nodes) connected to determine the functionality of the network (Demuth *et al.*, 2008). West (2000) carried out several experiments to measure the performance five different types of the neural network for credit scoring. While conducting experiments, West (2000) observed that Logistic regression is slightly more accurate in prediction in comparison to neural networks. This Research also noted that CART and k Nearest Neighbour results are not par with logistic regression. The neural network requires being trained on a dataset to predict the outcome of decision variables correctly (Presky *et al.*, 1996). Gately (1995) discussed applications of using the neural network in financials domains such as fraud detection in credit card transactions, forecasting company bankruptcy, classifying bad or good loan application and others areas where neural networks are successful (Ghosh and Reilly, 1994). Desai *et al.* (1996) compared the performance of a neural network and logistic regression and found that neural network able to correctly predict loan portfolio when the measure of success is accurately classifying bad loans only.

3.5 Geospatial

Geospatial data is a dataset which contains or provide information about geographical location/s. To analysis geospatial data, one requires a system that can interpret and process geographic data about latitude and longitude and assist decision makers in providing insights out of that data. Such systems are called Geographical Information System (Keenan, 1998). In recent years, we have seen rapid enhancement in the technology. As a result, the amount the spatial data available from satellite and user mobile data has been growing.

Can (1998) said that for housing and mortgage spatial data is a critical aspect as housing information remain as is in geographical space. In credit scoring system, one can combine spatial information of a particular location such as employment, property value, property area, average income, etc., with financial data to build a robust predictions model. Can (1998) noted that geospatial data is important for any business and policy, still its usability in mortgage and credit assessment is limited. In recent years, some researchers attempted to incorporate spatial data to estimate house prices (Tse, 2002), Carling and Lundberg (2005) combined the geographical information with loan data to examine the credit rationing.

Availability of high-end GIS software and fast computing environment makes it easier to utilise its power to strength credit scoring model along with the machine learning. By doing this not only bank and financial institutions to monitor or predict bad loans based on location, but also enable them to make new business strategies to reach out to uncovered audience or market.

3.6 Data Visualisation

Data volume has been increasing day by day, and it has become difficult to analyse the data at once using tables and reports. And it is known that human brain retains more information when it is received visually. Therefore, need for visual analytics has been increased from last few years and is proliferating. Data visualisation helps understanding complex data visually by easy pattern recognition, trends and provides granularity.

Data volume has been increasing day by day, and it has become difficult to analyse the data at once using tables and reports. And it is known that human

brain retains more information when it is received visually. Therefore, need for visual analytics has been increased from last few years and is proliferating. Data visualisation helps understanding complex data visually by easy pattern recognition, trends and provides granularity. Data visualisation also helps a user to play with data by making alterations. It also provides ease of improvement, classification of relevant factors that can enhance consumer behaviour, easily predict sales trends and customer behaviour.

Data visualisation tools such as Qlik, Tableau, R Shiny have played a significant role in demonstrating analytics and driving data insights to the users. Such tools are easy to operate compared to traditional statistical tools and software; that has led to enhancement in Business Intelligence. To explain results of advanced analytics and predictive algorithms to all users, it is essential to present the results to maintain performance visually.

Residential mortgages data consists of the geographical distribution of house locations. Sun *et al.* (2013) stated that data visualisation analyses and quickly derive stories efficiently and interactively. Organizations are extensively using data visualisation tools; as this software support drilling down the information and filtering the data as per requirement. Such software provides a facility of combining all the required information on a single platform called dashboard. Data visualisation supports Geo spatial data very well, and our work is primarily dependent on geographical locations of residences. This research work focuses on combinations of longitudes and latitudes that helps in identifying exact address of a house.

Because of the high volume of geospatial data, it is important to maintain latency between residential data and output generated by predictive models. For the reasons as mentioned above, data visualization is essential for our work that will help to visualize the results for the end users.

Chapter 4

Methodology

4.1 Overview

To assist financial auditor or stakeholder at financial institutions and banks, and to identify such loan portfolio which may default in future based on the geospatial information and financial data. This research work followed the KDD process which involves characteristics variables selection, perform data restructuring, data transformation and data mining for the deployment of a predictive model using visual analytics tools such as Tableau, QlikView, etc.

Software & Tools used:

Following is the list of tools and software that has been used while working on this project:

Data Processing: MS Excel 2017 and Alteryx Desginer 11.0

Version Control: Github (github.com)

Dashboard: Tableau Professional 10.2 and R Stuio 1.0.36

Data Storage: Github Pages (https://pages.github.com/) and Google Drive

R Packages used:

Packages required Logistic Regression Model: Following packages used to building simple regression and logistic regression based model for predicting the good or bad loan portfolio: glm() with class set to "binomial" for Logistic Regression and "log" for Poisson regression, ROSE, ROCR, Dplyr, maps, ggplot2

Decession Tree: Following r-packages used for building a predictive model based on decision tree: caret, rpart, rattle, ROSE, ROCR, RColor-Brewer, party, partykit

R Shiny: R Shiny packages for building interactive dashboards: leaflet, maps, ggmap, gridExtra, htmlwidgets, reshape2. To deploy predictive model on Tableau to build dynamic and easy to use dashboard R Server used

One may replicate our work on his/her computer having minimum hardware specifications outlined here. This research work carried on following machines.

Table 4.1: System configurations used to carry out this research

Specification	**System 1 - Lenovo Yoga 500**	**System 2 - Dell Inspiron 15**
Operating System	Windows 10 Professional	Windows 7 Professional
Processor	Intel(R) Core(TM) i3-5005CU @ 2.00GHz	Intel(R) Core(TM) i3-3217U @ 1.80GHz
RAM	4.00 GB	4.00 GB
System Type	64-bit OS, x64-Based Processor	32 -bit Operating System

4.2 Data Processing & Analysis

4.2.1 Overview

One requires the accessibility to the right set of data, and information on which statistical and modelling techniques can be applied to start any data oriented research in analytics domain, KPMG, Ireland provided data set. This data set contains historical data of various loan portfolios that maintained by each branch of banks or financial institutions. Also, this dataset has geospatial information about credit account along with their transactional history of previous loans. Credit scoring model requires being trained with a correct set of characteristics variables to provide the prediction with high accuracy.

This project has been carried out in four stages as outlined below:

- Data Selection & Processing
- Model Design & Implementation
- Testing & Model Results
- Deployment & Visualizations

4.2.2 Data Set

Dataset format: .xlsx
Number of attributes: 35
Total number of records: 237,390
All the variables and attributes have been carefully studied and analysed to decide what key factors will be used to develop the model. Based on the availability of RAM on the current system, it was decided to build a model on selected characteristics variables. One may train the model with all possible variables as well if system hardware allows. Below is the list of variables in original dataset:

```
[1] "ContractRef"            "LoanBalance"    "InterestType"
[4] "ProbationaryLoans"      "MortgageType"   "NewLoan"         "NIM"
[8] "DefaultedLoans"         "CreditRating"   "InterestIncome"  "LTV"
[12] "LTVCategory"           "MortgageYears"  "PropertyValue"
[15] "MaturityDate"          "BookingDate"    "LastValuationDate"
[18] "County"                "Branch"         "Address"
[21] "Town"                  "InArrears"      "AddressLongitude"
[24] "AddressLatitude"       "DaysInArrears"  "ArrearsCategory"
[27] "HousePriceMovement"    "ValueInArrears" "ValuationAgeYears"
[30] "UpdatedPropertyValue" "LTVUpdated"     "LTVCategoryUpdated"
[33] "CreditRatingMovement" "InterestRate"   "AnnualPYMT"
```

Below is the comprehensive list of all variables that have been chosen for the model creation:

ContractRef : Unique reference number assigned to each portfolio

InterestType : There are three types of interest rate: Fixed, Tracker and Variable

MortgageType : Whether property is bought for "buy-to-let" or "owner occpied"

NewLoan : Is portfolio is new or existing?

ProbationaryLoans : Has loan been taken on probation?

DefaultedLoans : Classify if the loan has defaulted in the past

LTVCategory : 5 Level categorized pre-assigned to each loan account

CreditRating : Each account is rated from 1-5 scale on the basis of credit union policy

MortgageYears : How many years mortgage has been taken for?

CreditRatingMovement : Percentage that indicates how credit rating has moved from previous value for an application

LTV : Ratio of applied loan amount to property evaluation value

LoanBalance : How much loan amount is left to repay?

InterestIncome : How much interest amount bank is earning?

PropertyValue : Recent property evaluation amount

AnnualPYMT : How much amount is getting repaid to the bank by the applicant annually?

AddressLatitude : Latitude value of the house on map

AddressLongitude : Longiitude value of the house on map

County : Name of the county where house is located

InArrears : Any amount that has not been paid earlier on time

ArrearsCategory : Category that defines duration of Arrears such as more than 90 days

Structure of the Data

```
Classes tbl_df, tbl and 'data.frame':    36696 obs. of  20 variables:
 $ ContractRef         : chr  "00000CONTR00111034" "00000CONTR00146183"
    "00000CONTR00175040" "00000CONTR00171901" ...
 $ InterestType        : Factor w/ 3 levels "Fixed","Tracker",..:
    2 3 2 1 2 3 2 2 2 3 ...
 $ MortgageType        : Factor w/ 2 levels "Buy to Let",
     "Owner Occupied": 1 2 2 2 2 2 2 2 2 2 ...
 $ NewLoan             : Factor w/ 2 levels "No","Yes":
    1 1 1 1 1 1 1 1 2 1 ...
 $ ProbationaryLoans   : Factor w/ 2 levels "No","Yes":
    2 1 1 1 1 1 1 1 1 1 ...
```

```
$ DefaultedLoans      : Factor w/ 2 levels "No","Yes":
   2 2 2 2 2 2 2 2 2 2 ...
$ LTVCategory         : Factor w/ 11 levels "> 100%","0 to 10%",..:
   11 8 11 5 3 5 9 11 9 9 ...
$ CreditRating        : Factor w/ 5 levels "1","2","3","4",..:
   4 2 4 4 3 2 3 3 4 2 ...
$ MortgageYears       : int  31 30 30 29 29 32 28 35 29 31 ...
$ CreditRatingMovement: int  3 0 0 0 2 -3 0 0 0 0 ...
$ LTV                 : num  0.983 0.65 0.93 0.368 0.167 ...
$ LoanBalance         : num [1:36696, 1] -0.647 -0.297 1.418
   -0.986 -1.32 ...
 ..- attr(*, "scaled:center")= num -1.05e-17
 ..- attr(*, "scaled:scale")= num 1
$ InterestIncome      : num [1:36696, 1] -0.71 -0.132 1.53
   -0.203 -1.1 ...
 ..- attr(*, "scaled:center")= num -2.14e-17
 ..- attr(*, "scaled:scale")= num 1
$ PropertyValue       : num [1:36696, 1] -1.3 -0.311 0.779
   -0.511 -0.205 ...
 ..- attr(*, "scaled:center")= num -2.51e-17
 ..- attr(*, "scaled:scale")= num 1
$ AnnualPYMT          : num [1:36696, 1] -1.2756 -0.2211
   0.9057 -0.3651 ...
 ..- attr(*, "scaled:center")= num 9.48e-18
 ..- attr(*, "scaled:scale")= num 1
$ AddressLatitude     : num  52.4 53.3 52.8 53.7 53.4 ...
$ AddressLongitude    : num  -7.7 -6.27 -6.74 -6.68 -6.21 ...
$ InArrears           : Factor w/ 2 levels "No","Yes":
   1 1 1 1 1 1 1 2 1 2 ...
$ County              : Factor w/ 26 levels "Carlow","Cavan",..:
     22 6 1 17 6 9 16 22 6 6 ...
```

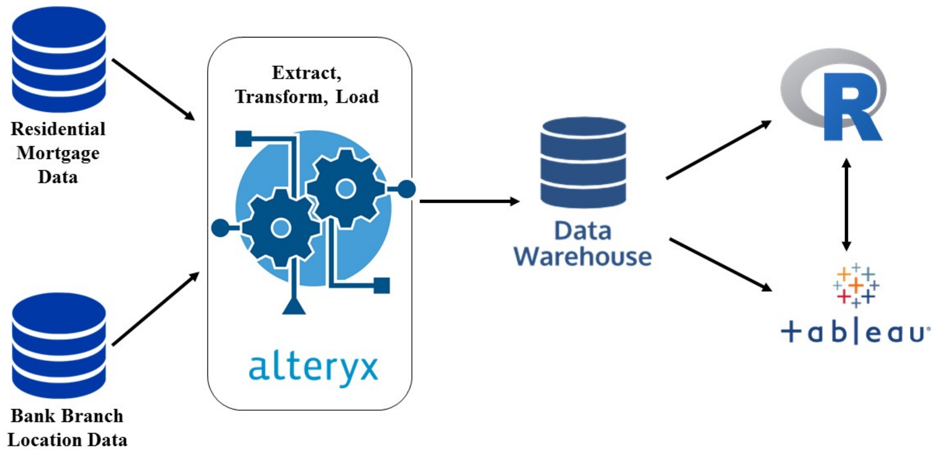


Figure 4.1: ETL & Data Model Architecture
Source: Designed using MS Office

```
$ ArrearsCategory     : chr  "0" "0" "0" "0" ...
```

4.3 Implementation

4.3.1 Data Extraction

Prior building the predictive model in R one, need to process and analyse the data. The primary objective is to identify any outliers and to normalise the available data set. Sola and Sevilla (1997) observed that un-normalised data tends to increase square mean error and then deviate the model prediction. Therefore, it is important to treat data and normalised it's all variables so that model works with high precision and accuracy. One can also do data pre-processing using R as well, but Alteryx provides graphical user interface for selecting features and settings that makes whole data processing phase easy and fast

Alteryx Desginer tool allow one to build workflow to prepare data from multiple data sources on the go and by using features such as 'Select', 'Random Sample', 'Transform' and 'Output' one can easily prepare data for the predictive model (Dinsmore, 2016). Alteryx can process a large amount of data set and optimised it to be ready for data modelling in R.

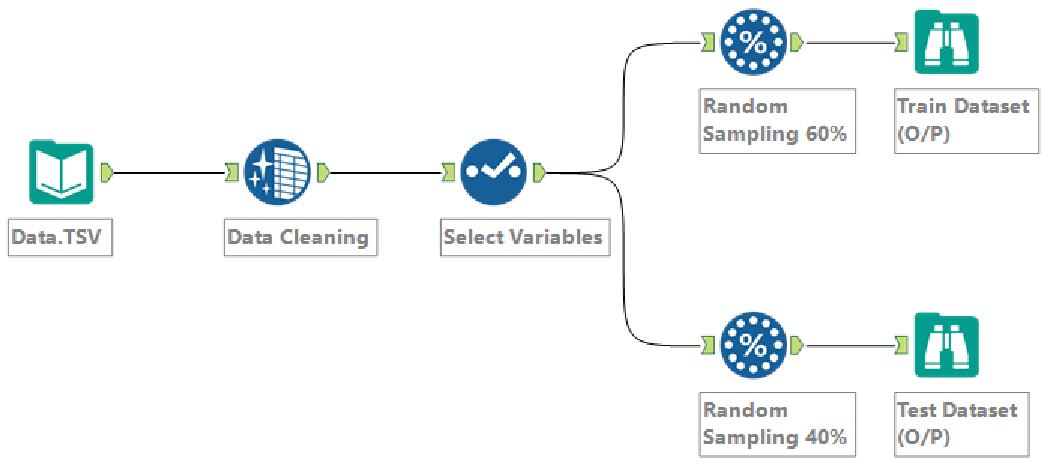


Figure 4.2: Data Processing using Alteryx
Source: Designed using in Alteryx Designer v11

In fig.4.2, raw data has been read using *Input tool*, then null values, white spaces etc removed using *Cleansing Tool* and variables selection has been done using *Select Tool*. To create train data set and test data set *Random Sample % tool*, which allows generating sample datasets.

4.3.2 Data Transformation

In Alteryx, there is no provision to normalize data. Processed data from Alteryx is loaded into **R Studio** for data normalization or scaling using in built functions such as scale($< variable >$) and log($< variable >$) on $LoanBalance$,

$PropertyValue$, $InterestIncome$ and $AnnualPYMT$ as these variables are crucial paramters for credit scoring to make unbaised prediction model.

R Studio: Data from Alteryx is loaded to R Studio for the development of prediction model. R is used to identify patterns or correlation in variables using *ggplot2*, *plot.ly*, *leaflets*. Two predictive models have developed based Logistic Regression and Decision Tree algorithms and both models performance evaluated concerning accuracy. Trained model is saved on the hard drive and loaded in Tableau, and with the help of R Server, Tableau allows the user to build dynamic visualizations. In Tableau, calculated fields can dynamically invokes R engine to perform calculations and then R results from output values back to Tableau, so that visualisations can be designed.

4.3.3 Data Loading

Integration of R in Tableau: Processed and transformed data is loaded into Tableau for building business dashboards. Credit analyst or auditors will use the dashboard to identify locations where the most number of loan default happenings or identify those portfolios which have provided incorrect information, etc. business decisions can be made with the help of credit scoring dashboard.

Installtion of R Server: Local instance of R Server is deployed by installing *Rserve* package from R console. To invoke R Server with following command:

```
install.packages("Rserve")
library(Rserve)
Rserve()
```

Setting in Tableau:

In Tableau, go to *Settings and Performance* under *Help* menu and then select *Manage External Service Connection*. Following settings are required to connect with R server:

```
Server: "localhost" or "127.0.0.1
Port: 6311
```

R scripts are written in calculated fields of Tableau to make calls to R using in built functions in Tableau such as *SCRIPT_STR* and *SCRIPT_REAL*

4.4 Predictive Model

4.4.1 Overview

Shmueli and Koppius (2011), define predictive analytics as the process of building statistical models using data mining algorithm with an objective to predict the outcome on future data set. A model is evaluated based on its predictive power or accuracy. As discussed in section 3.4, Logistic regression and Decision Tree are most commonly algorithms for building predictive models for credit scoring. Based on the requirement of predictive algorithms, data type of certain variables has been converted using below code:

```
Datav2$CreditRating <- as.factor(Datav2$CreditRating)
Datav2$InterestType <- as.factor(Datav2$InterestType)
Datav2$MortgageType <- as.factor(Datav2$MortgageType)
Datav2$NewLoan <- as.factor(Datav2$NewLoan)
Datav2$ProbationaryLoans <- as.factor(Datav2$ProbationaryLoans)
Datav2$LTVCategory <- as.factor(Datav2$LTVCategory)
Datav2$InArrears <- as.factor(Datav2$InArrears)
Datav2$County <- as.factor(Datav2$County)
Datav2$DefaultedLoans <- as.factor(Datav2$DefaultedLoans)
Datav2$LoanBalance <- scale(Datav2$LoanBalance)
```

```
Datav2$PropertyValue <- scale(Datav2$PropertyValue)
Datav2$InterestIncome <-scale(Datav2$InterestIncome)
Datav2$AnnualPYMT <-scale(Datav2$AnnualPYMT
```

4.4.2 Logistic Regression

Logistic regression is the most generally used technique for credit analysis and assessment as it works on binary response variables, i.e., 0 or 1 (Hilbe, 2011). In fig. 4.3, output results of standard logistics regression function lies between 0 and 1 only. In this research work output of response variable, i.e., the probability of default $p = 1$ is considered as 'Yes' and $p = 0$ is considered as 'No'. Probability is represented using logistic function (logit) and the probability of binary response variable based on the one, or more independent variables.

Model Settings:

Response Variable: DefaultedLoans
Family (Function): "Binomial" (Logit)

Model Implementation Details:

Initially, To train the model for all variables available in the dataset, but the model couldn't be trained because R engine failed to allocate 5.0GB vector space for the model. Following the line of code is used:

```
library(stats)
m2 <- glm(DefaultedLoans ~., family = "binomial", data = trainDatav2)
```

Next, model is trained with selective variables set and following code is used:

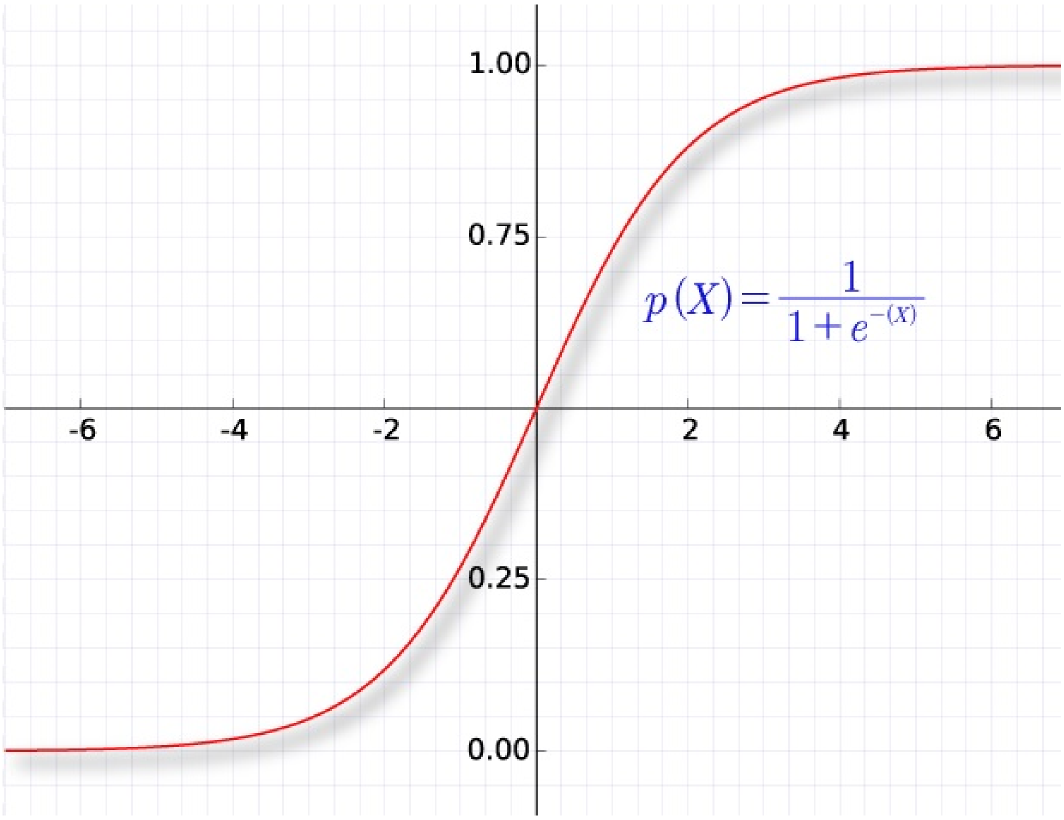


Figure 4.3: Standard Logistic Regression
Source: http://www.thefactmachine.com/wp-content/uploads/2015/03/13-Sigmoid.gif

```
simpleglmv2 <- glm(DefaultedLoans ~ CreditRating + InterestIncome +
    log(PropertyValue) + log(LoanBalance) + AnnualPYMT + LTV +
    InterestType + NewLoan + ProbationaryLoans + MortgageYears +
    MortgageType + InArrears + County + AddressLatitude + AddressLongitude,
    family = "binomial", data = trainDatav2)
```

Trained model is used to predict output for test dataset using following code:

```
testDatav2$prediction <- predict(simpleglmv2, newdata=testDatav2,
type="response")
```

4.4.3 Decission Tree

As discussed in section 3.4, Decision Trees has two most commonly used algorithm for credit scoring i.e. CART and C4.5. Classification and regression trees (CART) has been implemented using rpart() package available in R to build predicive model. rpart() syntax is

```
rpart(formula, data=, method=,control=)
```

formula = DefaultedLoans NewLoan + County + LoanBalance + PropertyValue + InterestIncome + CreditRating + AnnualPYMT + County + LTV + LTVCategory + InArrears + MortgageType + MortgageYears + AddressLatitude + AddressLongitude

data = trainDatav2

method = "Class"

control = Parameters for controlling the growth of tree.
control = rpart.control(minisplit=500,cp = 0.001)) At least 500 observations should be on a node before attempting a split and reduce the split fit factor by 0.001 before being attempted.

Packages such as rattle(), RColorBrewer(), etc. used to enhance the overall decision tree.

Model Implementation Details:

```
library(rpart)
library(rattle)
library(rpart.plot)
library(RColorBrewer)
library(party)
library(partykit)
library(caret)
```

```
defaultLoanTree <- rpart(DefaultedLoans ~ NewLoan + County + LoanBalance
+ PropertyValue + InterestIncome + CreditRating + AnnualPYMT + County
+ LTV + LTVCategory + InArrears + MortgageType + MortgageYears
+ AddressLatitude + AddressLongitude ,method = "class",data=trainDatav2,
control =  rpart.control(minisplit=5,cp = 0.001))

save(fit, file = "Model/classificationTreeV2.rda")
print(defaultLoanTree)
prp(defaultLoanTree)
tree.1 <- defaultLoanTree
fancyRpartPlot(tree.1)
```

Finally, the Model performance of logistic regression and decision tree has been evaluated based on GINI, ROC metrics.

4.5 Tableau & Dashboards

Tableau professional software is used to develop the business dashboard that will be utilised by end users such as credit analyst, auditors, banks officials, etc. In Tableau, CSV file connector is used to connect to the data source (sample dataset); then it is used to prepare various graphs and geospatial dashboard. The calculated field in Tableau allows making the call to R engine directly. By using calculated field options in Tableau, the predictive model is loaded into Tableau to make direct calls to R engine. Instructions and settings mentioned in section 4.3.3 used as is to connect Tableau with R.

In the dashboard, the user can select an origin city or region and distance (in miles) from that origin. Based on these inputs user will be able to take the business decision such as investigating a loan account when property value of a particular house is higher than the area average property value, or opening new branches near by to areas for which a significant number of loan applications is coming in. Following calculations are performed in Tableau calculated fields:

Calculation for the distance from Origin city:

```
3959 * ACOS
(
  SIN(RADIANS(LOOKUP(AVG([Address Latitude]), First()))) *
  SIN(RADIANS(AVG([Address Latitude]))
) +
  COS(RADIANS(LOOKUP(AVG([Address Latitude]), First()))) *
  COS(RADIANS(AVG([Address Latitude])))
  * COS(RADIANS(AVG([Address Longitude])) -
  RADIANS(LOOKUP(AVG([Address Longitude]),
  First())))
)
```

Calculation script for logistic regression model in Tableau:

```
SCRIPT_REAL('mydata <- data.frame(DefaultedLoans=.arg1, CreditRating=.arg2,
InterestIncome=.arg3, LoanBalance =.arg4, AnnualPYMT =.arg5, LTV =.arg6,
InterestType=.arg7,NewLoan=.arg8, ProbationaryLoans = .arg9,
MortgageYears=.arg10,MortgageType=.arg11, InArrears =.arg12,County =.arg13,
AddressLatitude=.arg14, AddressLongitude=.arg15, PropertyValue=.arg16);
load("Model/simpleglmv2.rda")

prob <- predict(simpleglmv2, newdata = mydata, type = "response")',
ATTR([Defaulted Loans]),ATTR([Credit Rating]),AVG([Interest Income]),
AVG([Loan Balance]),AVG([Annual PYMT]),AVG([LTV]),ATTR([Interest Type]),
ATTR([New Loan]),ATTR([Probationary Loans]),AVG([Mortgage Years]),
ATTR([Mortgage Type]),ATTR([In Arrears]),ATTR([County]),
AVG([Address Latitude]),AVG([Address Longitude]),AVG([Property Value]))
```

Chapter 5

Results

5.1 Overview

Model prediction accuracy of original test data set was 99.65%, which is practically impossible. As discussed in chapter 1 original data received from KPMG was made up using pre-defined formulas and rules to make it look real. Data didn't cover all possible scenario for a loan portfolio and achieving an accuracy of 99% in credit scoring model is difficult as one needs to train model recursively with large data size covering all permutations and combinations of situations for loan default.

5.2 Data Normalization

Original data set consist of 237389 observations and 35 variables, according to data 5% of loan applications have defaulted, and the customer has credit rating 1 will not default ever. Therefore, to consider all possible scenarios data has been normalised, and a data subset has been generated from original data set to carry experiments.

In Table 5.1, it is evident that data is very well structured and it does not give much information about the applicants who can default in future even if they had an excellent credit history. So loan portfolios of credit rating 1,2 and 3 don't contribute much to the objective of the project. Also, it can be seen that as per the distribution of original data, the only applicant with credit rating 5 will default. And this information does not comply to the real world scenarios. In discussion with KPMG, loan portfolios were again analysed to establish data that resembles the real world. Based on the variables such as loan balances, unemployment rates, annual income, address and mortgage years, changes had been made to the data that can be seen in Table 5.2. After data normalisation, all cases are considered including some of the extreme cases. Normalised data gives a better representation of loan portfolios accordance to the real world with 0.22% probability of default with credit rating 1, almost 11% defaulters with credit rating 4 and 10% chances of default and 1% chance of not default in credit rating 5.

Table 5.1: Distribution of Defaulted Loans vs Credit for Original Data

	Defaulted Loan?	
Credit Rating	**Yes**	**No**
1	44.93%	0%
2	21.24%	0.01%
3	17.40%	0.01%
4	10.96%	
5	0.89%	4.57%

5.3 Predictive Model Performance

Decision Tree over Logistic Regression:

Long *et al.* (1993) studied decision tree application for classifying heart disease patient and compared the performance of decision tree with logistic regression. Long *et al.* (1993), also noted that logistic regression model failed to consider missing data and decision tree model easily worked when data was

Table 5.2: Distribution of Defaulted Loans vs Credit for Normalized Data

	Defaulted Loan?	
Credit Rating	**Yes**	**No**
1	25.71%	0.22%
2	20.95%	0.79%
3	17.27%	1.41%
4	11.21%	10.78%
5	1.81%	9.86%

noisy. Satchidananda and Simha (2006), build credit scoring model and found that decision tree produce a more precise model and good performance in comparison to logistic regression.

Two individual predictive models were built on logistic regression and decision tree and both models performance on original data and normalised. Performance of the models has been compared using three key metrics AUROC, KS, Gini. AUROC is the area under receiver output characteristics, and an excellent model has AUROC score in the range of 80 - 90%. KS is Kolmogorov-Smirnov (KS) Goodness-of-Fit Test, and it is used to determine the classification power of the binary model, higher the score means better is classification power of a predictive model. Gini (or Gini index) is another more commonly used goodness of fit test in machine learning, and it has direct relation with AUROC, i.e. ($Gini = 2 * AUROCC - 1$).

In this research work, decision tree performance did better against the logistic regression performance. In Table table. 5.3 it can noted that decision tree AUROC (Area Under Receiver Output Characteristics) score is 81.4%, and logistic regression AUROC is 67.61%. Based on the results of previous research work and after considering current experiments results on normalised dataset, it is appropriate to build the business dashboard using decision tree model.

Another advantage of using decision tree model is that one can control the growth of decision tree using 'split' setting by doing so model performance can be optimised. On the other hand, to train model with logistic regression, one to select the restricted number of independent variables, otherwise, the model can not be trained with many variables as vector size response variable grows exponentially.

Significant Variables in Model:

```
[1] "CreditRating"                "PropertyValue"
[3] "LoanBalance"                 "LTV"
[5] "NewLoanYes"                  "ProbationaryLoansYes"
[7] "MortgageTypeOwner Occupied" "CountyCavan"
[9] "CountyCork"               "CountyDublin"
```

Table 5.3: Test results for Logistic Regression and Decision Tree performance

	Logistic Regression			**Decision Tree**		
Data/Measure	**AUROC**	**KS**	**Gini**	**AUROC**	**KS**	**Gini**
Original Data	99.82	15	10	99.72	99.38	99.44
Normalized Data	67.61	24	16	81.4	59.96	62.8

A model with higher AUROC on test data doesn't signify that the model is over-fitted, but it means that predictive has excellent performance. In this project, train data and test data created using original data set gave AUROC of 99% from which an inference has been noted that our model is over-fitted, but it may not always be a case.

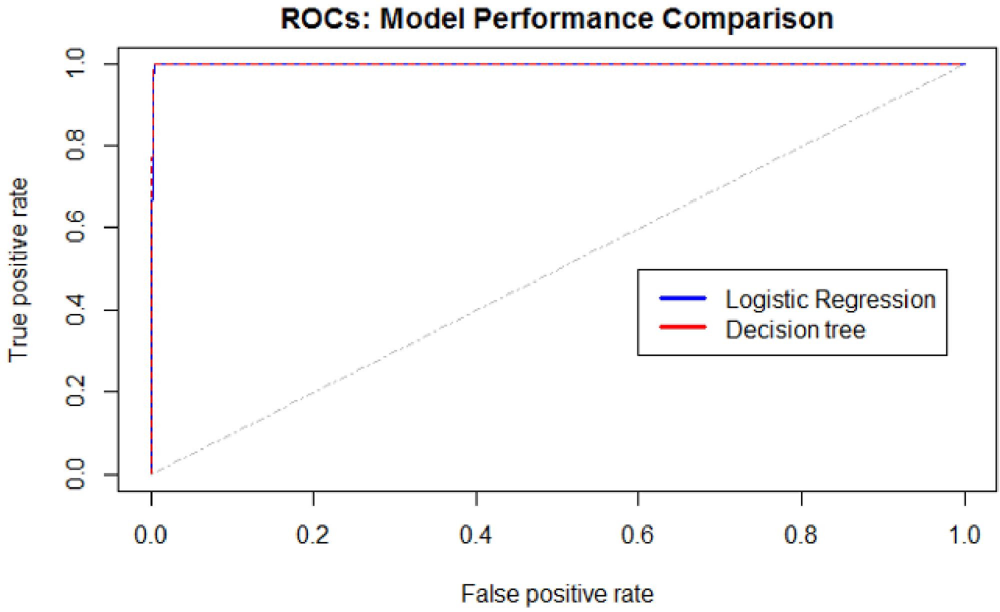


Figure 5.1: Original Data: ROCs for logistic regression vs decision tree
Source: Plotted in R Studio

Receiver operating characteristic (ROC) is one of technique to estimate the performance of the predictive model by plotting true positive rate(TPR) against the false positive rate(FPR). In figs. 5.2 & 5.1, performance of logistic regression and decision tree has been compared for ROC index. The original dataset has a 90-degree line for both logistic regression and decision tree, which suggests that predictive might be overfitted. It is evident from the ROC for the normalised data set in fig. I5.2 that decision tree performance is better over logistic regression for credit assessment and analysis. In general, higher the area under of ROC (AUROC) curve signifies better performance.

The decision tree of original data set is represented in fig.**??**, it can be seen

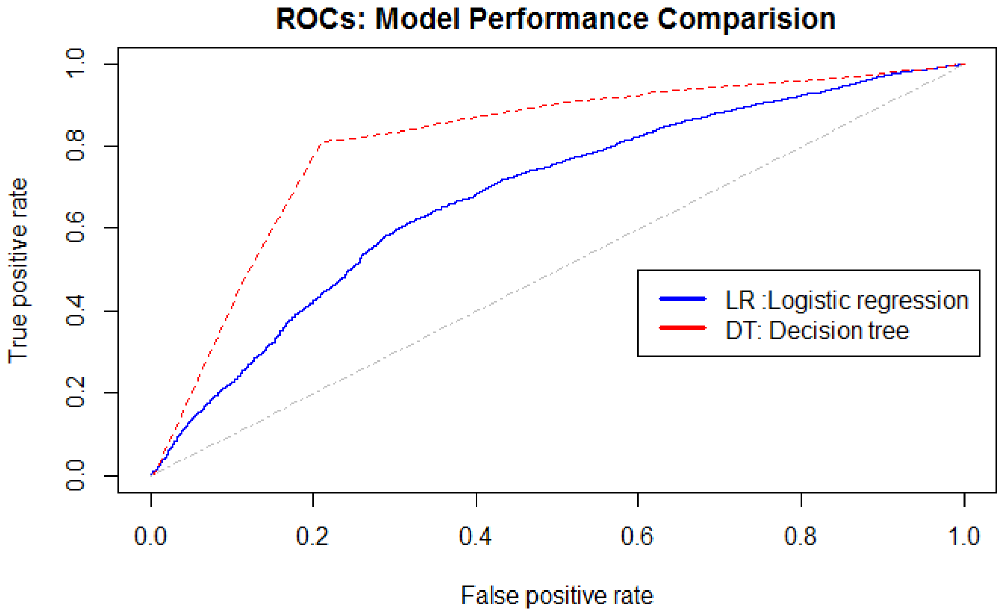


Figure 5.2: Normalized Data: ROCs for logistic regression vs decision tree
Source: Plotted in R Studio

that 72% observations have been classified based on one rule, i.e. CreditRating ¡=4.5. Due to this reason original data set was giving the accuracy of 99

In fig. 5.4, the decision tree for normalised data set is represented, and this tree has over 2000 rules which further improves its performance. Due to limitations of page size, we are showing a decision tree with a limited number of nodes. The originally trained tree has 328 nodes, and depth of tree was 9.

5.4 Tableau Dashboard

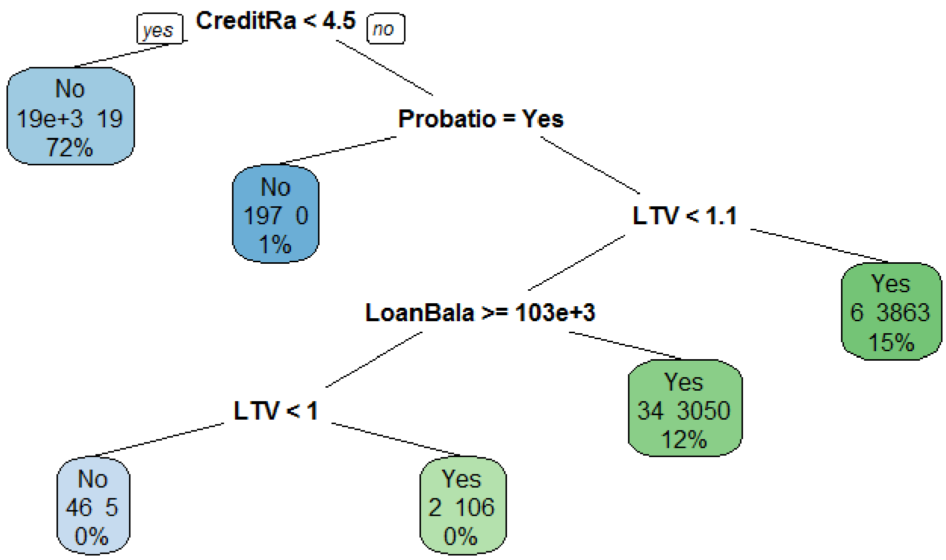


Figure 5.3: Decision tree of original data
Source: R Package:(Milborrow, 2016)

5.4.1 Overview

Considering the day-to-day requirement of stakeholders at KPMG, three business dashboards are designed using Tableau software. Primary purpose of building the dashboard using Tableau, as it allows the easy integrations with a predictive model from R engine and provides a better means to visualise prediction and forecast. This dashboard offers an interactive way to identify outlier and analyse loan accounts and the user can drill down to street level analyses as shown in fig.**??**. Also, it integrates multiple data sources that turn out to be a great time saving for an auditor, as it is not needed to physically map variables from different sources such as employment rate of a town with listed portfolios.

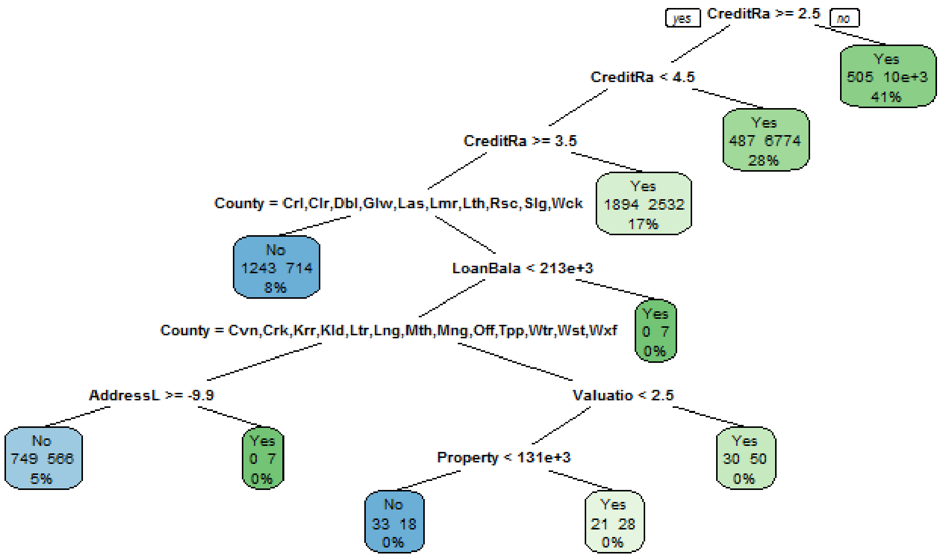


Figure 5.4: Decision tree of modified data

Source: R Package: (Milborrow, 2016)

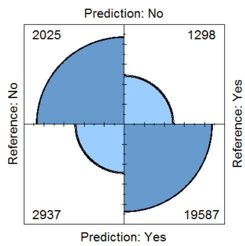


Figure 5.5: Confusion Matrix of Test Data

Source: R Studio

5.4.2 Predictive Dashboard

To support end user decision-making process, probability of default is segmented into five levels 0-5%, 6-25%,26-6%,61-80%, and 81-100%. A heat map is shown in fig 5.8, this will assist credit analyst to find out those loan account which requires particular human attention.

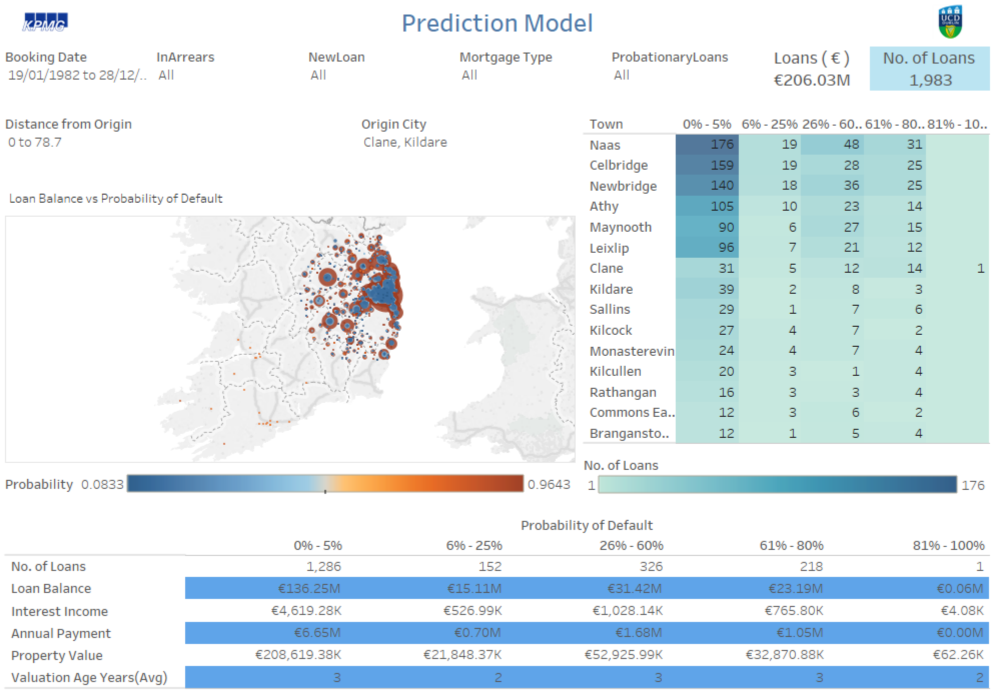


Figure 5.6: Predictive Model Dashboard

Source: Tableau Professional v10

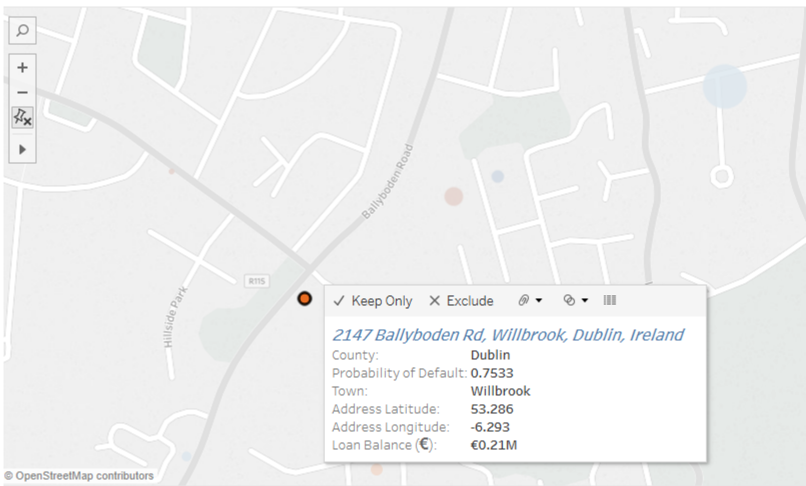


Figure 5.7: Street view map analysis

Source: Tableau Professional v10

Town	0% - 5%	6% - 25%	26% - 60..	61% - 80..	81% - 10.
Dublin	155	16	30	24	1
Lucan	78	7	19	7	
Swords	60	4	11	7	
Blackrock	51	2	13	5	
Dublin 4	39	3	14	4	
Dublin 6W	41		13	5	
Crumlin	36	5	14	3	
Malahide	31	5	11	8	
Rathfarnham	38	5	8	3	
Clontarf East	29	7	5	4	
Drumcondra	25	4	4	9	
Cabra East	34		3	2	
Dublin 6	28	2	6	3	
Balbriggan	28	1	4	4	
Merchants Q..	18	4	10	5	

Figure 5.8: Probability Heat map
Source: Tableau Professional v10

Case 1: Division of dashboard into three parts turned out to be a good idea as it provides better prediction results: Loan Balance vs Probability of Default, Top fifteen towns and statistical summary table based on selection filter and values. It can be seen from fig. 5.9, a number of loans booked during Irish property crash (2007-2010) have the highest probability of default.

Case 2: With the use of action filters, dashboard provide detailed information about a number of loan account and loan balance due from selected origin city. As in fig. 5.10, there is only one credit account which satisfies selected user criteria. $3, Woodlawn, MalahideDemesne, Malahide, Co.Dublin, Ireland$ probability of default is 0.15%, and remaining loan balance is $0.07M$.

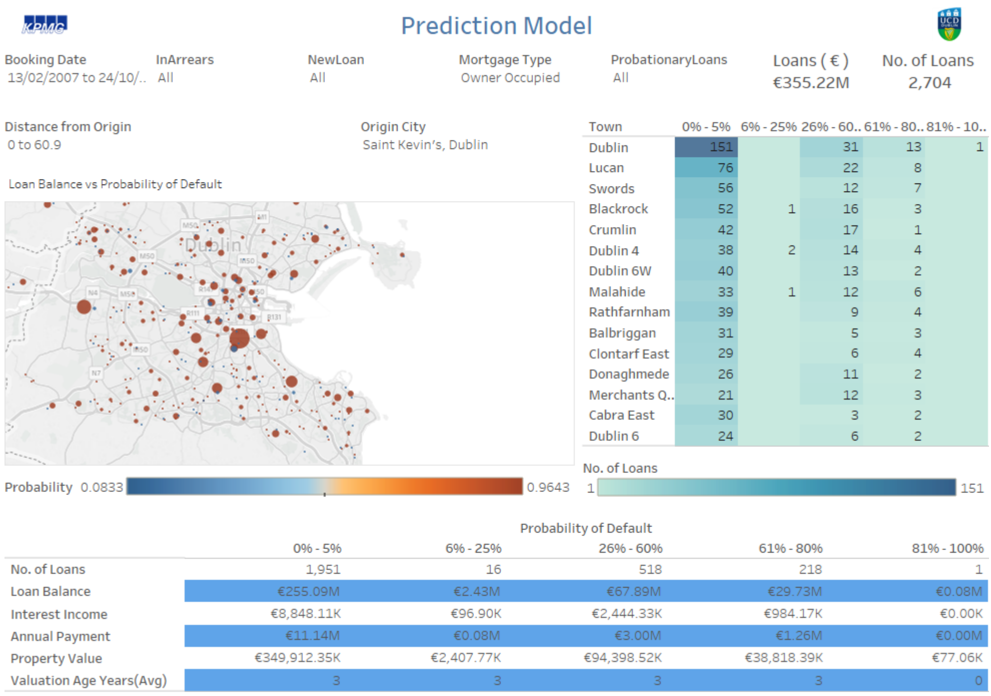


Figure 5.9: Irish Property Crash analysis

Source: Tableau Professional v10

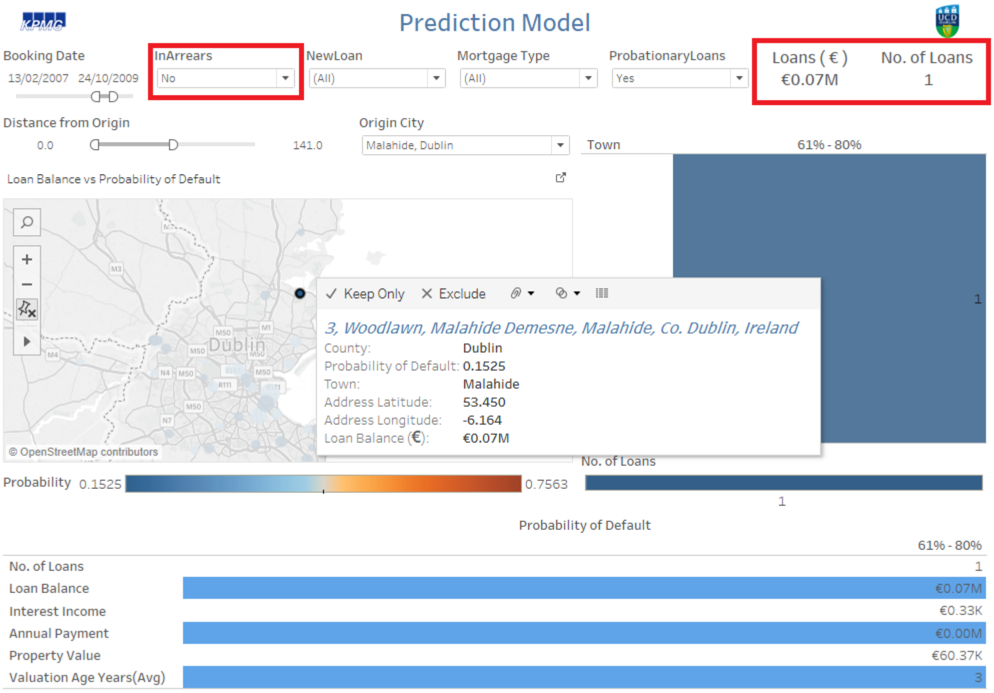


Figure 5.10: Street view map analysis

Source: Tableau Professional v10

Chapter 6

Discussion

6.1 Introduction

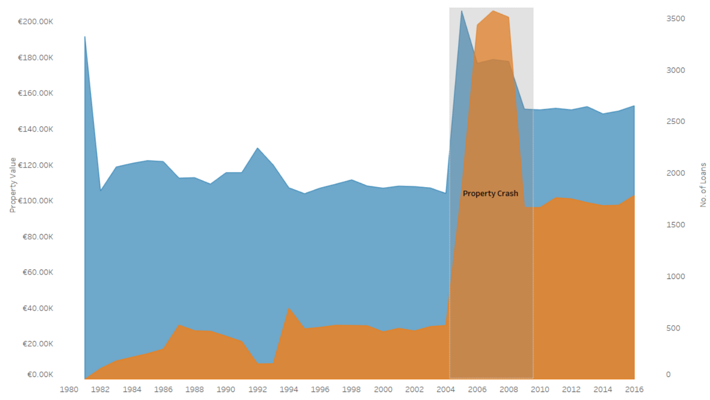


Figure 6.1: Irish Property Crash
Source: Tableau Professional v10

This chapter presents the detailed discussion and analysis of patterns and trends discovered during this research work. Keeping the interest of every stakeholder from bank officer to auditors, an attempt has been build simplicity in the business dashboard so that end user can use it efficiently to drive the business decision. When it was discovered that original data is not appropriate from the predictive modelling perspective, the modified data has been used throughout this research work. All analysis has been presented considering modified data set.

6.2 Patterns & Analysis

Since the property crash (2007-2010) average property price and the number of loan applications has been reduced as it can be seen in fig 6.1. Earlier the average property was €120K then it increased by 100%.

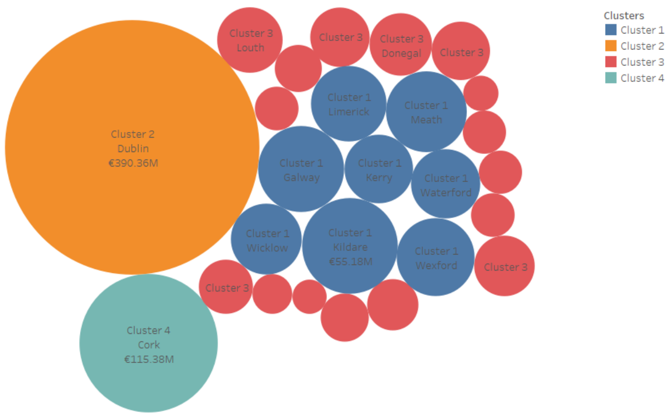


Figure 6.2: County Cluster
Source: Tableau Professional v10

The k-mean algorithm is used to cluster twenty-six(26) counties among four clusters based on the loan balance, as Ireland property market vary a lot in a county. Co. Dublin is classified into cluster #2, and Co. Cork is classified into cluster #4 has highest outstanding loan balance. As shown in fig. 6.3 8 counties are in cluster #1 and 16 counties with least outstanding loan balance classified into cluster #3.

Inputs for Clustering

Variables: Sum of Loan Balance
Level of Detail: County
Scaling: Normalized

Summary Diagnostics

Number of Clusters: 4
Number of Points: 26
Between-group Sum of Squares: 0.94145
Within-group Sum of Squares: 0.0075545
Total Sum of Squares: 0.949

Clusters	Number of Items	Centers Sum of Loan Balance
Cluster 1	8	3.7193e+07
Cluster 2	1	3.9036e+08
Cluster 3	16	1.5238e+07
Cluster 4	1	1.1538e+08
Not Clustered	0	

Figure 6.3: Clustering Results
Source: Tableau Professional v10

During data analysis phase, it was noted that the most properties have LTV(Loan-to-value) ratio between 60 - 80%. There are few properties with LTV higher than 100%. In fig. 6.4

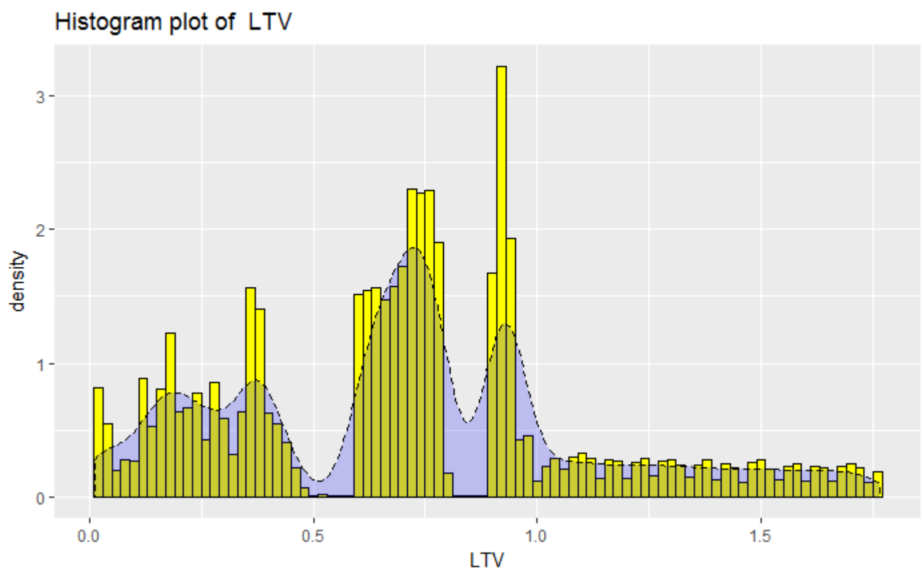


Figure 6.4: Loan to Value histogram
Source: R Studio ggplot()

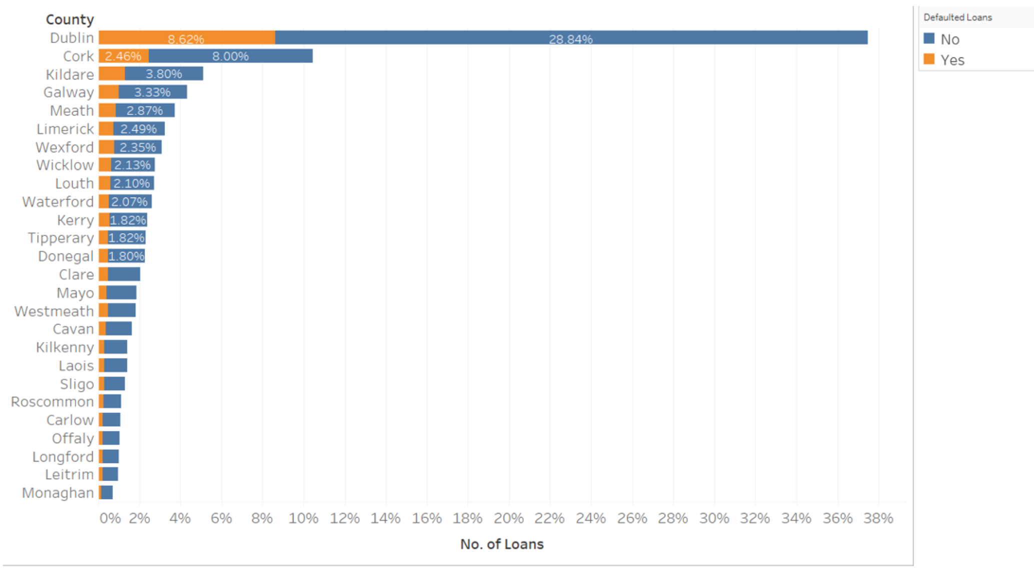


Figure 6.5: Number of account county wise

Source: Tableau Pro

A total number of accounts in the normalised data set was 273,000, and most numbers of default and loan account were from County Dublin and Cork as seen in fig. 6.5.

6.3 Statistical Analysis

Descriptive analysis has been performed using SPSS tool to get a better understanding of data and take required steps for data preprocessing. It can be seen from fig.6.6 and fig.6.7, there are no missing values or unknown values present in the data as it is very well structured. Statistical analysis helped in generating the rules to normalise data for building an unbiased predictive model. *LoanBalance* is of 6 digits (e.g. €103,584), and *AnnualPayment* is of 4 digits (e.g. €4614); in such scenarios, a predictive model can assign more weight to *LoanBalance* as it is more significant compared to the *AnnualPayment*. So, the model will outweigh *LoanBalance* over *AnnualPayment* but for credit assessment both the variables hold equal importance. Therefore, its is necessary to scale the variables in data set accordingly to achieve better performance.

Statistics

		MortgageYears	CreditRatingMovement	LTV	LoanBalance	InterestIncome
N	Valid	237389	237389	237389	237389	237389
	Missing	0	0	0	0	0
Mean		31.51	-.18	.6533	$103,584.1347	$3,548.3352
Median		32.00	.00	.7100	$102,743.1748	$2,637.5155
Std. Deviation		2.290	1.441	.28922	$60,944.43737	$3,175.87151
Range		7	8	1.75	$447,521.24	$30,189.24
Minimum		28	-4	.01	$2,316.76	$0.00
Maximum		35	4	1.77	$449,838.00	$30,189.24

Figure 6.6: Analysis of Significant variables - 1

Source: SPSS

Statistics

		PropertyValue	AnnualPYMT	AddressLatitude	AddressLongitude
N	Valid	237389	237389	237389	237389
	Missing	0	0	0	0
Mean		$157,246.4406	$5,014.7292	53.1237	-7.3283
Median		$144,962.7367	$4,614.6601	53.2926	-6.9333
Std. Deviation		$68,430.28540	$2,205.19901	.69122	1.13915
Range		$797,233.39	$26,808.28	3.94	4.45
Minimum		$50,184.43	$1,438.98	51.43	-10.46
Maximum		$847,417.82	$28,247.26	55.38	-6.02

Figure 6.7: Analysis of Significant variables - 2

Source: SPSS

6.4 Dashboard

Considering the day-to-day requirement of stakeholders at KPMG, three business dashboards are designed using Tableau software. Main response to build the dashboard using Tableau, as it allows the easy integrations with the predictive model from R engine and provides a better means to visualise prediction and forecast.

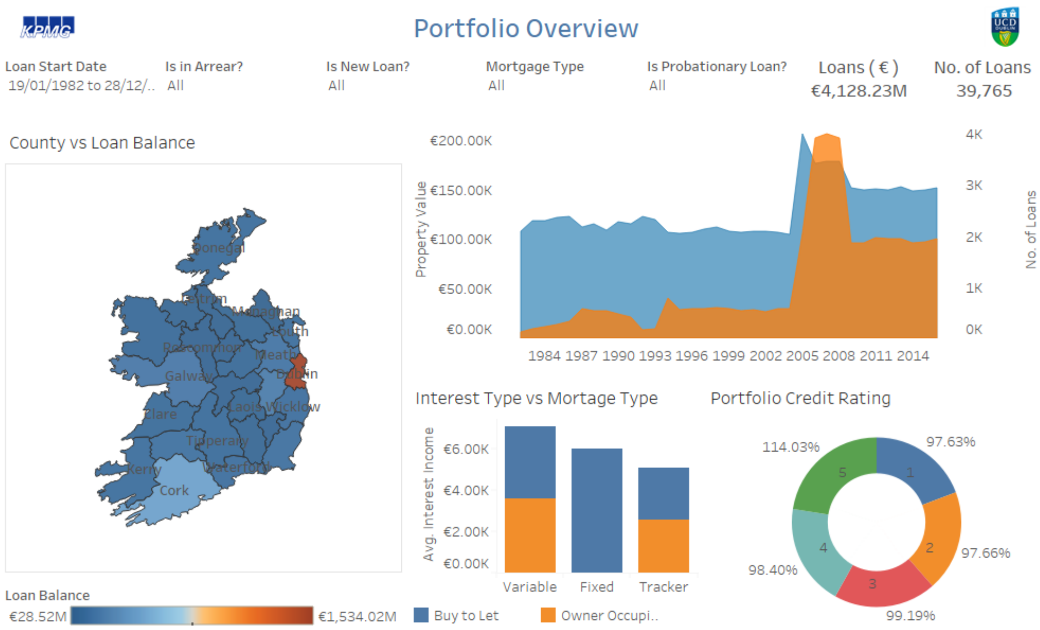


Figure 6.8: Portfolio Overview

Source: Tableau Professional v10

Portfolio Overview

Portfolio overview dashboard is built to allow auditors and credit analysts to analyse overall existing loan portfolio registered under a bank. This dashboard allows the user to perform a detailed analysis by selecting various combinations of variables from filters such as:

- Loan Start Date: User can view a selective number of loan account based on the start date of an account
- Is in Arrears?: Does a loan account has any outstanding repayment in last one year?

- Mortgage Type: For what purpose mortgage has been buy-to-let or owner occupied?
- Is probationary Loan?: Has the loan account been converted to probationary loan

The geospatial map allows the analyst to view loan balance for each county, along with a time-series analysis of property price and a number of accounts for past three decades. Using interest type vs mortgage type user can identify which type of interest is giving more income to banks.

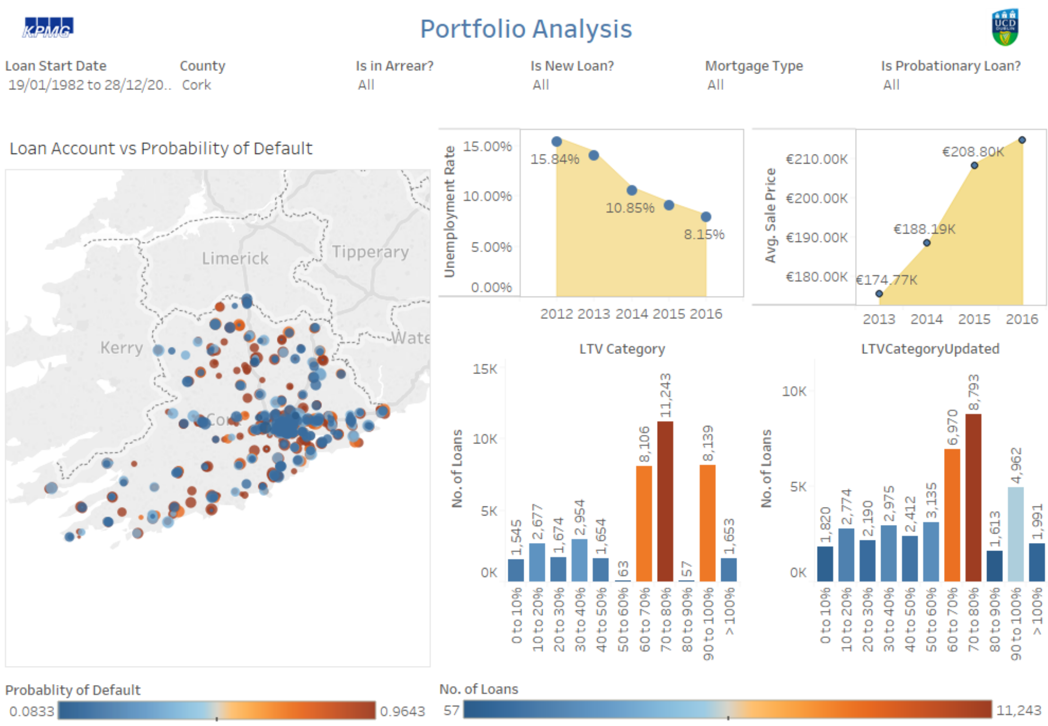


Figure 6.9: Portfolio Analysis

Source: Tableau Professional v10

Portfolio Analysis

Portfolio analysis dashboard allow user to view loan account movement from one LTV (loan to value) category to other, along with unemployment rate and average property sale price in a town. User can select range of property from map using a distance(radius in kms as in fig 6.10) to compare trends in selected neighbourhood.

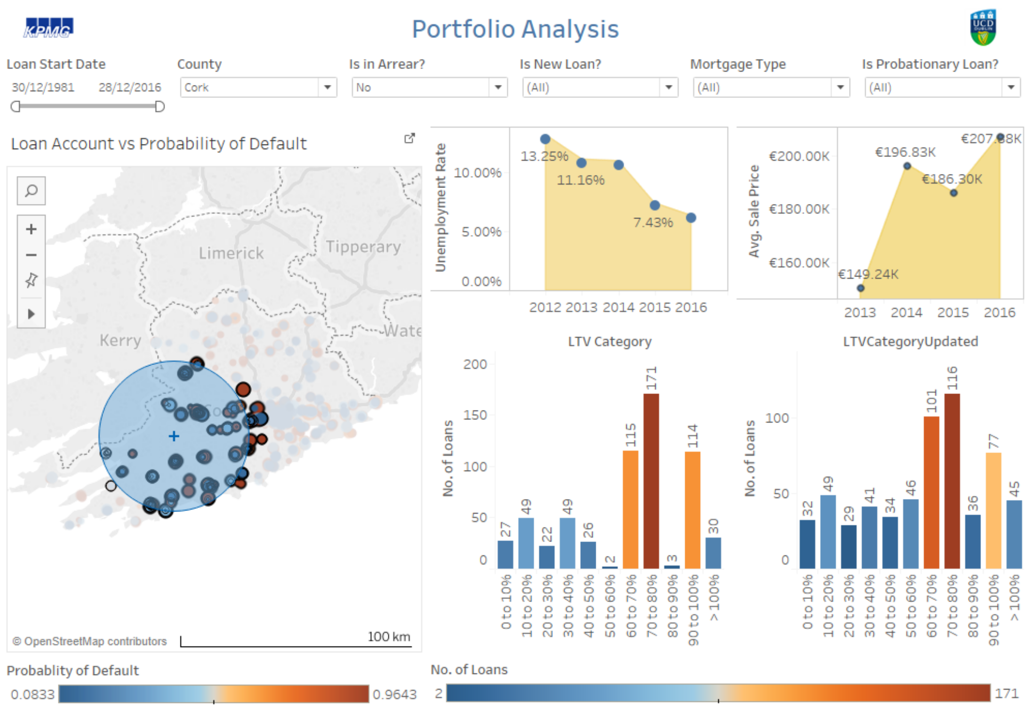


Figure 6.10: Property Selection using distance

Source: Tableau Professional v10

R Shiny Dashboard The second dashboard has been built using R shiny package, which allows one to create interactive dashboards. R shiny is quite a popular package among data scientist as it is open source and easy to deploy. However, only limitations it has that user won't be able to load large datasets, and during this research work, it was observed that dashboard performance was slow data set has more than 100,000 row. A dynamic dashboard has been build which allows user to view property based on clustering and it allows user can view street level statistical metrics as seen in fig 6.11.

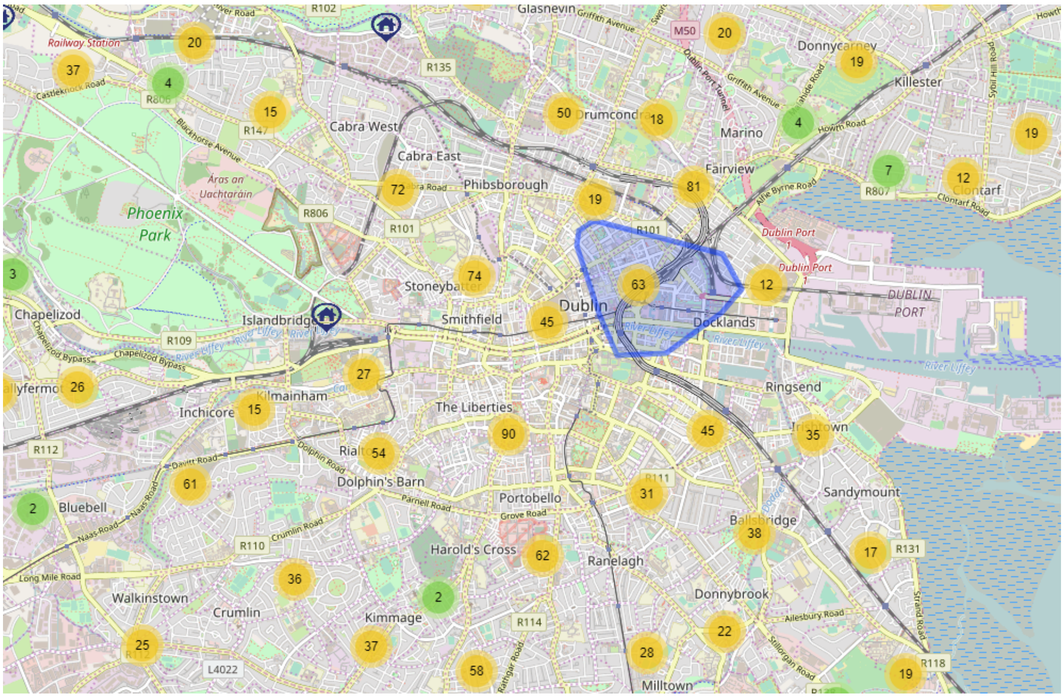


Figure 6.11: Dashboard Build using R Shinny

Source: R Studio

6.5 Success

To measure the performance of this work, a working business dashboard has been given to stakeholders to perform analysis and provide their feedback. The questionnaire covers all qualitative and quantitative measures which focus on the performance of the model, characteristics of a good visualisation, financial aspects and legal information. Based on the feedback received and suggestions dashboard features are improved to satisfy the user needs. The user was given three dashboard designs and asked to recommend best one with possible changes stating the reason for choice made and suggestions to further improve usability.

Following represents the overall feedback received from users:

- **47%**Auditors feel this dashboard is handy
- **70%**Feel that Geospatial techniques have enhanced credit assessment
- **60%**Likely to recommend this dashboard to colleagues
- **65%**Feel that dashboard is very appealing

Auditors, Credit Analysts, Data Scientists and Tableau Consultants have responded to the survey by providing unbiased and specific suggestions on all three dashboards. From the recorded responses, it is significant that dashboards are helpful for credit analysis or for making new bank sales strategies. Fig.6.1 represents overall positive or negative responses to the feedback.

Table 6.1: Overall Recorded Response

#	Answer	%
1	Extremely positive	37.50%
2	Moderately positive	50.00%
3	Slightly positive	12.50%
4	Neither positive nor negative	0.00%
5	Slightly negative	0.00%
6	Moderately negative	0.00%
7	Extremely negative	0.00%
	Total	100%

Chapter 7

Conclusions and Future Research

7.1 Future Work

Credit scoring is a sensitive and critical component of any bank and financial institutions. The outcome of this work presents a predictive model that connects with a business dashboard. One can integrate the day to day financial data from a bank with this dashboard. This work can be improvised with the help of real banking data so that predictive model can be trained efficiently. The dashboard is designed in a way which allows easy integration with any data.

Dashboard: Real transactional data can improve the performance of Tableau dashboard which acts as a decision support system

Geospatial Data: Credit assessment can be enhanced if it includes information such as house coordinates, neighbourhood amenities

External Factors: The model can perform better when trained with large number of external factors such as medical information, average salary in neighbourhood, inflation rates

7.1.1 Proposed system Architecture

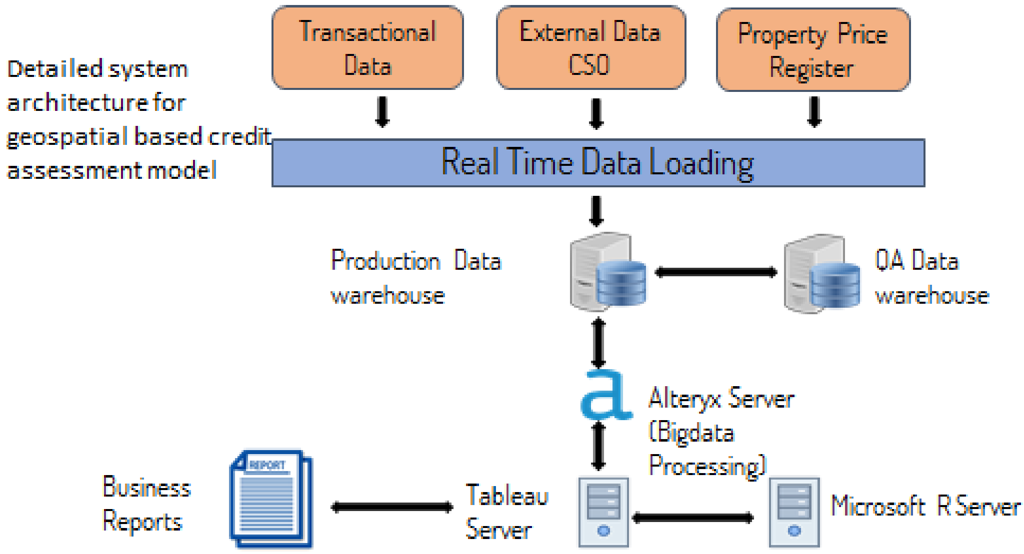


Figure 7.1: System Architecture

Source: MS Powerpoint

This practicum proposes an enhanced and scalable system architecture that can be used on a commercial platform (fig. 7.1). A high-end data warehouse can be used to process and store big data in real time by using techniques such as MapReduce, Hadoop clusters, etc. Deployment of Microsoft R Server will allow faster and efficient processing of predictive models; that eventually will generate dynamic reports on Tableau Server. Using such type of system architecture, a business can have multiple users who can access predictive dashboard; which is hosted on a central data centre; from various locations on their hand-held devices such as mobile phones, tablets, iPads. Banks or Financial Services Institutions can reduce manual efforts and dependencies from a monotonous task and complex procedures. Also, data from various sources will be available in a pre processed form from which user can take business decisions much faster than traditional methods.

7.2 Conclusion

The objective of this project was to build an interactive and efficient dashboard that supports credit analysis and assessment of residential loan portfolios with the help of Geospatial methods. This practicum was stemmed on aggregation of loan portfolio data and financial services data that adheres to credit assessment policies and macroeconomic performance indicators. The purpose of developing predictive models for calculating the probability of default using logistic regression and decision trees was successfully achieved. Initial review of the literature revealed that majority of the researchers believe models developed using logistic regression show better performance compared to decision trees, but few have concluded the opposite. This practicum falls into the category of those researchers, who have stated decision trees give better performance than logistic regression based on KS, GINI and ROC statistics. Although, because of the limitation of data, it cannot be said that the developed predictive model will show similar results when connected to real life dataset. There are possibilities that logistic regression can give better performance compared to decision trees. Also, if the model is integrated with another dataset, some training may be required to obtain properly fitted models.

Program Code

Program code along with data set uploaded on Github repository.

https://github.com/01dkg/Credit-Scoring

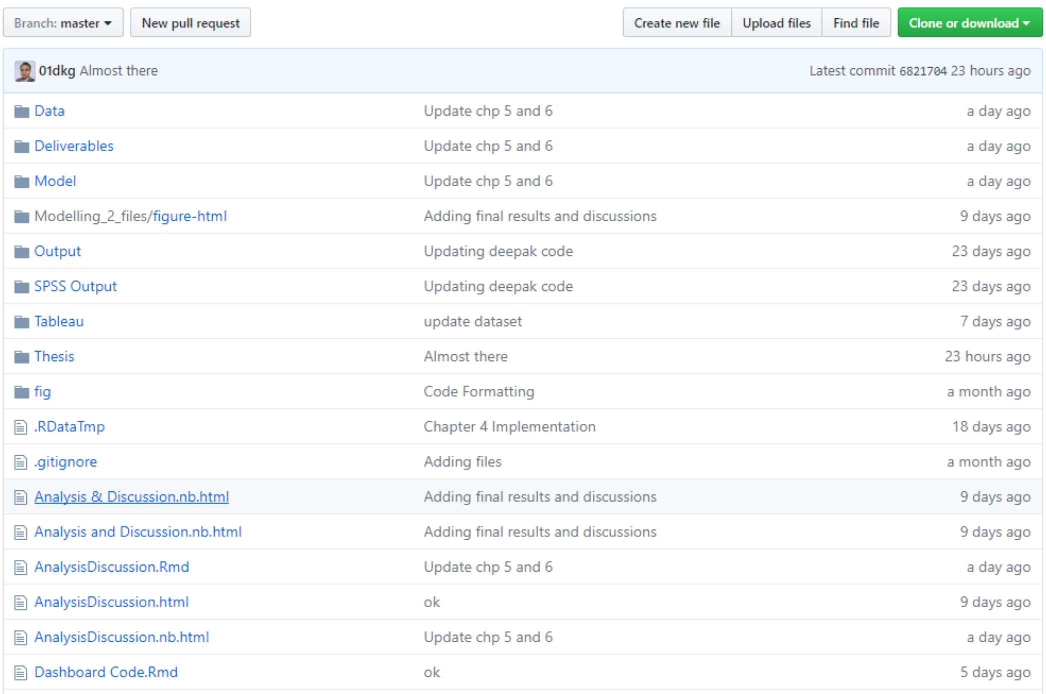


Figure 7.2: Git Repository

Descriptive Analytics

Frequency Table

Frequency Table - InterestType - July 26, 2017

InterestType

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Fixed	45349	19.1	19.1	19.1
	Tracker	118744	50.0	50.0	69.1
	Variable	73296	30.9	30.9	100.0
	Total	237389	100.0	100.0	

Frequency Table

Frequency Table - MortgageType - July 26, 2017

MortgageType

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Buy to Let	52085	21.9	21.9	21.9
	Owner Occupied	185304	78.1	78.1	100.0
	Total	237389	100.0	100.0	

Frequency Table

Frequency Table - LTVCategory - July 26, 2017

LTVCategory

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	> 100%	9802	4.1	4.1	4.1
	0 to 10%	9244	3.9	3.9	8.0
	10 to 20%	16112	6.8	6.8	14.8
	20 to 30%	9994	4.2	4.2	19.0
	30 to 40%	17628	7.4	7.4	26.4
	40 to 50%	9744	4.1	4.1	30.6
	50 to 60%	356	.1	.1	30.7
	60 to 70%	48483	20.4	20.4	51.1
	70 to 80%	67253	28.3	28.3	79.5
	80 to 90%	270	.1	.1	79.6
	90 to 100%	48503	20.4	20.4	100.0
	Total	237389	100.0	100.0	

Figure 7.3: Frequency table for Interest Type, Mortgage Type, LTV Category

Frequency Table

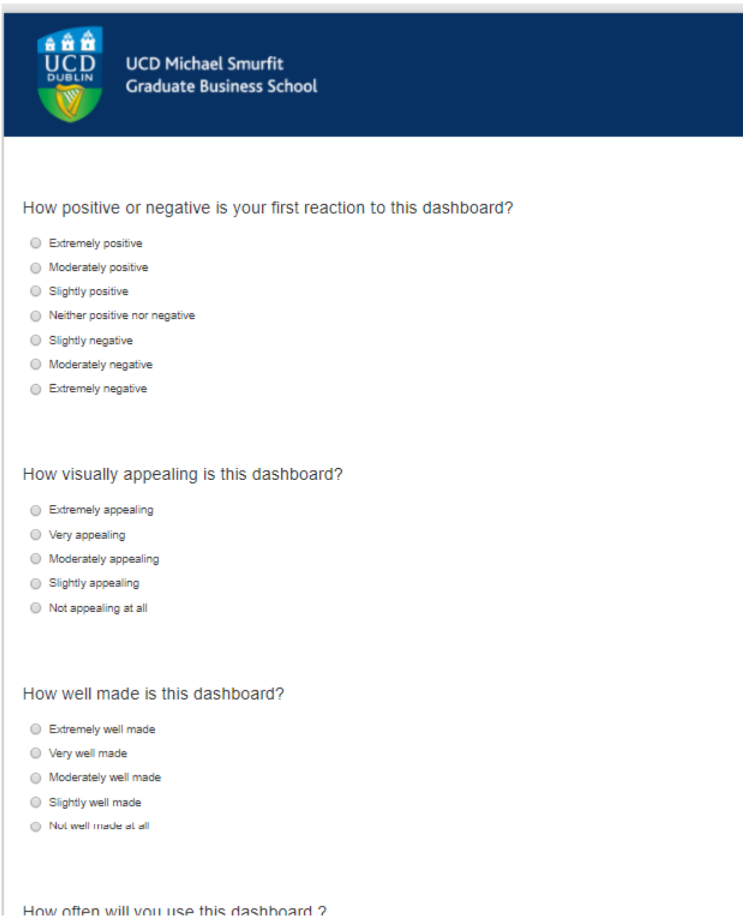
Frequency Table - County - July 26, 2017

County

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	Carlow	2697	1.1	1.1	1.1
	Cavan	3836	1.6	1.6	2.8
	Clare	5447	2.3	2.3	5.0
	Cork	26133	11.0	11.0	16.1
	Donegal	6608	2.8	2.8	18.8
	Dublin	77892	32.8	32.8	51.7
	Galway	12870	5.4	5.4	57.1
	Kerry	6813	2.9	2.9	59.9
	Kildare	10806	4.6	4.6	64.5
	Kilkenny	3606	1.5	1.5	66.0
	Laois	3403	1.4	1.4	67.4
	Leitrim	2377	1.0	1.0	68.4
	Limerick	8698	3.7	3.7	72.1
	Longford	2210	.9	.9	73.0
	Louth	6097	2.6	2.6	75.6
	Mayo	5617	2.4	2.4	78.0
	Meath	8282	3.5	3.5	81.5
	Monaghan	1788	.8	.8	82.2
	Offaly	2809	1.2	1.2	83.4
	Roscommon	3478	1.5	1.5	84.9
	Sligo	3540	1.5	1.5	86.4
	Tipperary	6060	2.6	2.6	88.9
	Waterford	6238	2.6	2.6	91.5
	Westmeath	4815	2.0	2.0	93.6
	Wexford	8024	3.4	3.4	96.9
	Wicklow	7245	3.1	3.1	100.0
	Total	237389	100.0	100.0	

Figure 7.4: Frequency table for County

Questionnaire Form



UCD Michael Smurfit
Graduate Business School

How positive or negative is your first reaction to this dashboard?

- Extremely positive
- Moderately positive
- Slightly positive
- Neither positive nor negative
- Slightly negative
- Moderately negative
- Extremely negative

How visually appealing is this dashboard?

- Extremely appealing
- Very appealing
- Moderately appealing
- Slightly appealing
- Not appealing at all

How well made is this dashboard?

- Extremely well made
- Very well made
- Moderately well made
- Slightly well made
- Not well made at all

How often will you use this dashboard ?

- Moderately useful
- Slightly useful
- Not useful at all

Do you think using Geo-spatial techniques, assessment of credit has improved?

- No
- Yes
- Other

How well do you think Geo-spatial techniques has enhanced credit assessment and analysis?

Not at all likely										Extremely likely
0	1	2	3	4	5	6	7	8	9	10

How likely would you be to recommend this dashboard to a colleague?

Not at all likely										Extremely likely
0	1	2	3	4	5	6	7	8	9	10

What do you like most about this dashboard?

What about this dashboard could be improved?

Bibliography

Abdou, H., J. Pointon and A. El-Masry. 2008. Neural nets versus conventional techniques in credit scoring in egyptian banking. *Expert Systems with Applications*, **35**(3): 1275–1292.

Abdou, H. A. H. 2009. *Credit scoring models for Egyptian banks: neural nets and genetic programming versus conventional techniques*. Ph.D. thesis, University of Plymouth.

Al Amari, A. 2002. *The credit evaluation process and the role of credit scoring: a case study of Qatar*. Ph.D. thesis, University College Dublin.

Altland, H. W. 1999. Regression analysis: statistical modeling of a response variable.

Altman, E. I. 1968. Financial ratios, discriminant analysis and the prediction of corporate bankruptcy. *The journal of finance*, **23**(4): 589–609.

Anderson, R. 2007. *The credit scoring toolkit: theory and practice for retail credit risk management and decision automation*. Oxford University Press.

Bailey, M. 2004. *Consumer credit quality: underwriting, scoring, fraud prevention and collections*. White Box Publishing.

Bensic, M., N. Sarlija and M. Zekic-Susac. 2005. Modelling small-business credit scoring by using logistic regression, neural networks and decision trees. *Intelligent Systems in Accounting, Finance and Management*, **13**(3): 133–150.

Beynon, M. J. 2005. Optimizing object classification under ambiguity/ignorance: application to the credit rating problem. *Intelligent Systems in Accounting, Finance and Management*, **13**(2): 113–130.

Breiman, L., J. Friedman, C. J. Stone and R. A. Olshen. 1984. *Classification and regression trees*. CRC press.

Can, A. 1998. Gis and spatial analysis of housing and mortgage markets. *Journal of Housing Research*, **9**(1): 61–86.

Capon, N. 1982. Credit scoring systems: A critical analysis. *The Journal of Marketing*, pages 82–91.

Carling, K. and S. Lundberg. 2005. Asymmetric information and distance: an empirical assessment of geographical credit rationing. *Journal of Economics and Business*, **57**(1): 39–59.

Crook, J. 1996. Credit scoring: An overview. *WORKING PAPER-UNIVERSITY OF EDINBURGH DEPARTMENT OF BUSINESS STUDIES*.

Demuth, H., M. Beale and M. Hagan. 2008. Neural network toolbox 6. *Users guide*, pages 37–55.

Desai, V. S., J. N. Crook and G. A. Overstreet. 1996. A comparison of neural networks and linear scoring models in the credit union environment. *European Journal of Operational Research*, **95**(1): 24–37.

Dinsmore, T. W. 2016. Self-service analytics. In: *Disruptive Analytics*, pages 199–230. Springer.

Durand, D. *et al.*. 1941. Risk elements in consumer instalment financing. *NBER Books*.

Eisenbeis, R. A. 1978. Problems in applying discriminant analysis in credit scoring models. *Journal of Banking & Finance*, **2**(3): 205–219.

Fisher, R. A. 1936. The use of multiple measurements in taxonomic problems. *Annals of human genetics*, **7**(2): 179–188.

Gately, E. 1995. *Neural networks for financial forecasting.* John Wiley & Sons, Inc.

Ghosh, S. and D. L. Reilly, 1994. Credit card fraud detection with a neural-network. In: *System Sciences, 1994. Proceedings of the Twenty-Seventh Hawaii International Conference on*, volume 3, pages 621–630. IEEE.

Gup, B. E. and J. W. Kolari. 2005. *Commercial banking: The management of risk.* John Wiley & Sons Incorporated.

Hand, D. J. and W. E. Henley. 1997. Statistical classification methods in consumer credit scoring: A review. *Journal of the Royal Statistical Society. Series A (Statistics in Society)*, **160**(3): 523–541. ISSN 09641998, 1467985X.
URL `http://www.jstor.org/stable/2983268`

Hand, D. J. and S. Jacka. 1998. Consumer credit and statistics. *Statistics in finance*, pages 69–81.

Hilbe, J. M. 2011. Logistic regression. In: *International Encyclopedia of Statistical Science*, pages 755–758. Springer.

Hosmer, D. W., B. Jovanovic and S. Lemeshow. 1989. Best subsets logistic regression. *Biometrics*, pages 1265–1270.

Huang, C.-L., M.-C. Chen and C.-J. Wang. 2007. Credit scoring with a data mining approach based on support vector machines. *Expert systems with applications*, **33**(4): 847–856.

Joanes, D. N. 1993. Reject inference applied to logistic regression for credit scoring. *IMA Journal of Management Mathematics*, **5**(1): 35–43.

Keenan, P. B. 1998. Spatial decision support systems for vehicle routing. *Decision Support Systems*, **22**(1): 65–71.

Koh, H. C., W. C. Tan and C. P. Goh. 2015. A two-step method to construct credit scoring models with data mining techniques. *International Journal of Business and Information*, **1**(1).

Kohavi, R. and J. R. Quinlan, 2002. Data mining tasks and methods: Classification: decision-tree discovery. In: *Handbook of data mining and knowledge discovery*, pages 267–276. Oxford University Press, Inc.

Koutanaei, F. N., H. Sajedi and M. Khanbabaei. 2015. A hybrid data mining model of feature selection algorithms and ensemble learning classifiers for credit scoring. *Journal of Retailing and Consumer Services*, **27**: 11–23.

Lee, T.-S. and I.-F. Chen. 2005. A two-stage hybrid credit scoring model using artificial neural networks and multivariate adaptive regression splines. *Expert Systems with Applications*, **28**(4): 743–752.

Liberati, C., F. Camillo and G. Saporta. 2017. Advances in credit scoring: combining performance and interpretation in kernel discriminant analysis. *Advances in Data Analysis and Classification*, **11**(1): 121–138.

Long, W. J., J. L. Griffith, H. P. Selker and R. B. D'agostino. 1993. A comparison of logistic regression to decision-tree induction in a medical domain. *Computers and Biomedical Research*, **26**(1): 74–97.

Milborrow, S. 2016. *rpart.plot: Plot rpart Models. An Enhanced Version of plot.rpart.* R package.
URL `http://CRAN.R-project.org/package=rpart.plot`

Nie, G., W. Rowe, L. Zhang, Y. Tian and Y. Shi. 2011. Credit card churn forecasting by logistic regression and decision tree. *Expert Systems with Applications*, **38**(12): 15273–15285.

Presky, D. H., H. Yang, L. J. Minetti, A. O. Chua, N. Nabavi, C.-Y. Wu, M. K. Gately and U. Gubler. 1996. A functional interleukin 12 receptor complex is composed of two β-type cytokine receptor subunits. *Proceedings of the National Academy of Sciences*, **93**(24): 14002–14007.

Satchidananda, S. and J. B. Simha. 2006. Comparing decision trees with logistic regression for credit risk analysis. *International Institute of Information Technology, Bangalore, India.*

Shmueli, G. and O. R. Koppius. 2011. Predictive analytics in information systems research. *Mis Quarterly*, pages 553–572.

Sola, J. and J. Sevilla. 1997. Importance of input data normalization for the application of neural networks to complex industrial problems. *IEEE Transactions on Nuclear Science*, **44**(3): 1464–1468.

Sullivan, A. 1981. Consumer finance. *EI Altman, Financial Handbook (9.3-9.27), New York: John Wiley & Sons.*

Sun, G., R. Liang, F. Wu and H. Qu. 2013. A web-based visual analytics system for real estate data. *Science China Information Sciences*, **56**(5): 1–13.

Thomas, L. C., D. B. Edelman and J. N. Crook. 2002. *Credit scoring and its applications*. SIAM.

Tse, R. Y. 2002. Estimating neighbourhood effects in house prices: towards a new hedonic model approach. *Urban studies*, **39**(7): 1165–1180.

West, D. 2000. Neural network credit scoring models. *Computers & Operations Research*, **27**(11): 1131–1152.

Xia, Y., C. Liu, Y. Li and N. Liu. 2017. A boosted decision tree approach using bayesian hyper-parameter optimization for credit scoring. *Expert Systems with Applications*, **78**: 225–241.

Zekic-Susac, M., N. Sarlija and M. Bensic, 2004. Small business credit scoring: a comparison of logistic regression, neural network, and decision tree models. In: *Information Technology Interfaces, 2004. 26th International Conference on*, pages 265–270. IEEE.

Zhang, D., X. Zhou, S. C. Leung and J. Zheng. 2010. Vertical bagging decision trees model for credit scoring. *Expert Systems with Applications*, **37**(12): 7838–7843.

Zhou, X., L. Yang and H. Hu. 2016. Research of thunderstorm warning system based on credit scoring model. In: *Frontier Computing*, pages 65–76. Springer.

Zhou, X., D. Zhang and Y. Jiang. 2008. A new credit scoring method based on rough sets and decision tree. *Advances in Knowledge Discovery and Data Mining*, pages 1081–1089.