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

List of Algorithms

Preface

> *Men occasionally stumble over the truth, but most of them pick themselves up and hurry off as if nothing had happened.*
>
> — Winston Churchill

Much of the front matter is optional. In particular, include things like a Dedication, List of Figures, List of Tables, List of Algorithms, only if there are enough of them to justify it and it would help the reader.

Don't include both a Foreword and a Preface since they perform similar roles.

The same goes for appendices, index, glossary, list of notation and terms at the end. Include if they would help the reader.

But always, of course, include the Bibliography.

University College Dublin
August 7, 2017

Xxxxx Xxxxxxx
Yyy Yyyyyyyyy

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/Executive Summary

Chapter 1

Introduction

Here's an example of a quote.

> If anybody calls, says the brother, tell them I'm above in Merrion Square workin at the quateernyuns, says he, and take any message. There does be other lads in the same house doing sums with the brother. The brother does be teachin them sums. He does be puttin them right about the sums and the quateernyuns.
>
> *Indeed.*
>
> I do believe the brother's makin a good thing out of the sums and the quateernyuns. Your men couldn't offer him less than five bob an hour and I'm certain sure he gets his tea thrown in.
>
> *That is a desirable perquisite.*
>
> Because do you know, the brother won't starve. The brother looks after Number Wan. Matteradamn what he's at, it has to stop when the grubsteaks is on the table. The brother's very particular about that.
>
> *Your relative is versed in the science of living.*
>
> Begob the sums and the quateernyuns is quickly shoved aside when the alarm for grub is sounded and all hands is piped to the table. The brother thinks there's a time for everything.

— Myles na Gopaleen, *Cruiskeen Lawn*

1.1 Background

We begin by ...

1.1.1 Particular Background

Here are some examples of indexing: Newton's algorithm is still widely used, with modifications.

Note that the `\index{algorithm!Newton}` gives an index subentry for Newton under the entry for algorithm. The index entries and/or their page numbers can be formatted using a pipe | symbol in the `\index{}` command as follows:

Definition 1.1. The strictest definition of an *algorithm* is: a finite set of instructions that can be carried out in a finite amount of time: that is, it must terminate.

These instructions must be clear and unambiguous as they are to be interpreted by a (dumb) machine, so we must be absolutely precise about their meaning — mathematical logic is thus crucial in the design of algorithms.

In practice, many useful numerical "algorithms" that we study may get closer and closer to the desired solution without reaching it in a finite time. So, typically, we accept as an "algorithm" a finite set of instructions that will get within any desired tolerance of the true solution in a finite time. If the algorithm is stochastic (involves probability, as many modern ones do) the term "metaheuristic" is sometimes used.

In particular, you could use the `\index{algorithm@\textit{algorithm}}` or `\index{algorithm|\textbf}` to indicate the first or most important occurrence in the text of the term "algorithm", etc.

Some minor examples of other things indexing can do:

- You can handle accented words as in école: the index entry appears in the correct order under E, as desired;
- You can put in cross-references, as in

  Are metaheuristics really algorithms?

Note: when you LaTeX your file `myfile.tex`, a file `myfile.idx` is produced by `\makeindex`; this file must be sorted by an operating system command, e.g.,

```
makeindex myfile
```

This generated a *sorted* index file `myfile.ind`. Running LaTeX one more time gets the index printed in the right place by `\printindex`.

Here is a dummy theorem to show how to reference notation:

Theorem 1.2. *Let $\mathbb{F}_q$ be a finite field of q elements. Then q is a power of some prime number p.*

1.2 Referencing

Recall the strictures against plagiarism. Accidental plagiarism is still plagiarism. If you paraphrase, you must still cite. If your paraphrase is very similar to the original, then delete it and quote instead (and cite).

Use a reference format similar to that used in the journal Management Science. This can be achieved by using the Management Science Endnote style or by using a style based on the Chicago 15th B style in Endnote. Please ensure that volume (and issue numbers where appropriate) are displayed, as well as appropriate page numbers.

The following are examples of suitable output:

Keenan (2003) identified the role of GIS...

Or GIS can seen as a form of IS (Keenan, 2003) ...

Do not put the title of the paper you are citing, normally.

Do not write: (Keenan, 2003) found that...

To insert a citation, use the `\cite` command in LaTeX, or `\citep` and `\citen` etc. if you know `natbib`.

Chapter 2

Business Background

2.1 Introduction

We begin this chapter by The credit industry has experienced two decades of rapid growth with signicant increases in installment credit, singlefamily mortgages, autoenhnancing, and credit card debt. Credit scoring models have been widely used by the financial industry during this time to improve cash flow and credit collectionsThe advantages of credit scoring include reducing the cost of credit analysis, enabling faster credit decisions, closer monitoring of existing accounts, and prioritizing collections

Chapter 3

Literature Review

3.1 Introduction

In recent years, purchasing capability of an **economy** has increased due to improvement in their finances, and employment levels. Ranging from buying small household items to **expensive items such as a house, a car or an office**. To buy a house or a car, one needs to have a large amount of money available to him; that is not necessarily possible most of the time.

Start with an exmaple to explain what is loan. There are certain critical circumstances that can occur anytime, where one may need a certain amount of cash. So one may need 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 near future. Lent amount is called principal amount and amount to be repaid is 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 taken in instalments Partial Payments]. A loan can be provided by banks, corporations and financial institutions. Banks and financial institutions provide 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 very large and banks cannot provide the loan based on verbal agreement, 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 in order to free his/her mortgage. In case, if an applicant is not able to repay the loan within 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 different 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.

Put **Photo of loan application process folw chart**

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 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 of time.

And, the major question is how do banks decide whether to give a loan or not? This question is of major 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, analyze 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 criticized because data that has been used consists of some assumptions to evolve model statistically.

3.2.2 Advantages and Disadvantages of Credit Scoring

In (Crook, 1996), Crook 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 behavior. A significant advantage of this approach is its reusability; 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 standardized 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 standardized 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?

(Al Amari, 2002) 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.

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 instituions making efforts to enhace tranditional credit scoring mechanisms by incorporating latest technology and

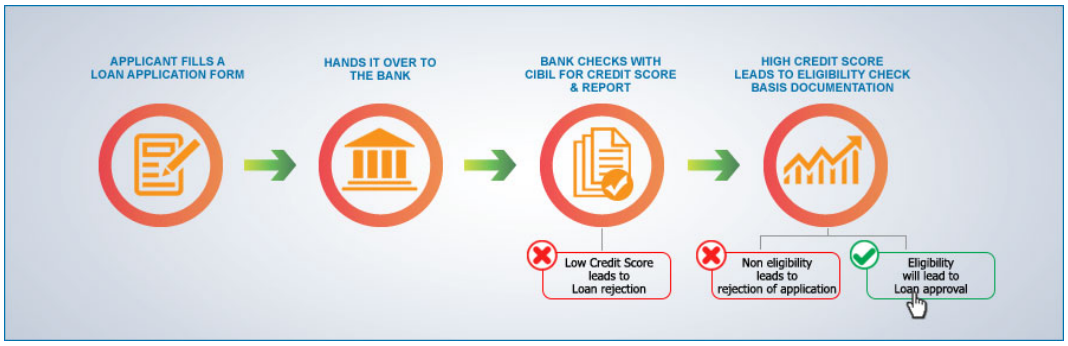
tools. Not only avaiablity data about customer but also rapid development in machine learning and analytics providing a foundation stone to banks.

Traditional credit scoring process with random selection of good and bad portfolio from creditors file around 50 - 300 Capon (1982) chartartestics points from loan portfolios to build a essential subset to perform statastical analysis. In (Hand and Henley, 1997), Hand mentioned about three commonly used approaches used for selecting characterstics out of avialble data: Expert Knowledge, Stepwise Statstical procedure and evaulating individual characterstics. Subject Matter Expert(SME)

Credit analysis and assessment is very important for banks and financial instituions to evaluate the credit worthiness of an applicant or a borrower. Banks implements various factors while assessing credit risk; such as credit rating, loan to value ratio, probability of default, etc.; that leads to 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 charterstics points. Thereafter credit score is used depending on the usage of 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 instituations incorportae two stage cut-offs, in this system if credit score is higher than upper cut-off then credit is granted straighted and denied if score is lower thant lower cut-off. If 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 higher than upper cut-off then credit is granted else denied.

Banks and financial instution sets their own cut-off for credit score based on the probablities of each applicant ability to repay or nonpayment of credit amount.



Adding flow chart of Evaulation Process and Pricing

However, Credit Risk has recevied a lot of critisim as well from Academics and Researchers. Al Amari (2002) has questioned about optimal method to evaulate customers? What are key variables or data points which an analyst must consider while evaulating 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).

Logistic Regression has resemblance with Linear regression and it is also most commonly used statsical technique for building credit scoring system. Dichtomous nature of logistic regression outcome probablity (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 techniqeue in machine learning and

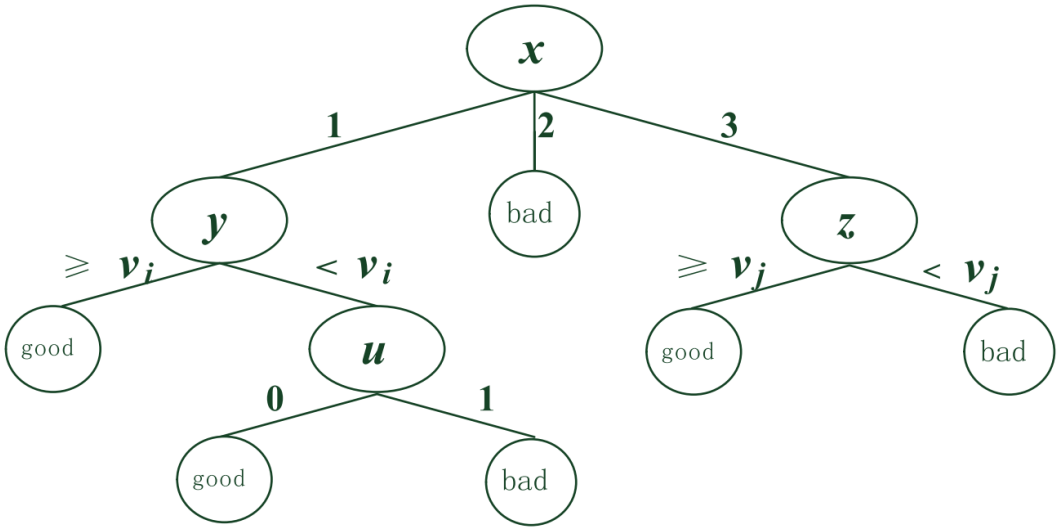


Figure 3.1: Simple Decision Tree Source: (Zhang *et al.*, 2010)

widely using for buidling credit scoring system. Classification & Regression Trees (CART) and C4.5 are two widely use decision tree algorithms (Kohavi and Quinlan, 2002). One of the firts 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 misclassifications(Zekic-Susac *et al.*, 2004). This process of selecting 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). In 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 stateliite and user mobile data has been growing.

In (Can, 1998), Can 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.citepcan1998gis, also 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.

Availablity 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 growing rapidly. 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 analyze 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 growing rapidly. Data visualization helps understanding complex data visually by easy pattern recognition, trends and provides granularity. Data visualization 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 behavior, easily predict sales trends and customer behavior.

Data visualization 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 re-

sults 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 visualization analyses and quickly derive stories efficiently and interactively. Organizations are extensively using data visualization 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 visualization supports Geo spatial data very well, and our work is primarily dependent on geographical locations of residences. Our 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.

4.2 Software & Tools Specifications

Following is the list of tools and softwares 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 Logisctic Regression Model: Following packages used to building simple regression and logistic regression based model for predictig the good or bad loan portfolio: glm() with class set to "bionomial" 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, RColorBrewer, 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 dyanmic and easy to user dashboard R Server used

4.3 Hardware Specifications

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

4.4 Data

4.4.1 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.

4.4.2 Data Dictionary

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. Below is the comprehensive list of all variables that has 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

4.5 Implementation

One can install packeges in r studio using common install.packages(¡packagename¿)

what is r packages, how to use r packege, which are essential

This packages is used for following purpose

4.6 Deployment & Connection Setup

Chapter 5

Results

5.1 Introduction

The results ...

Chapter 6

Discussion

6.1 Introduction

In this chapter we examine ...

Chapter 7

Conclusions and Future Research

– That's a most foolhardy remark, he said sharply, because the nerve-strings and the sheep's head itself are whirling into the same bargain and you can cancel out one whirl against the other and there you are — like simplifying a division sum when you have fives above and below the bar.

– To say the truth I did not think of that.

– Mollycules is a very intricate theorem and can be worked out with algebra but you would want to take it by degrees with rulers and cosines and familiar other instruments and then at the wind-up not believe what you had proved at all. If that happened you would have to go over it till you got a place where you could believe your own facts and figures as exactly delineated from Hall and Knight's Algebra and then go on again from that particular place till you had the whole pancake properly believed and not have bits of it half-believed or a doubt in your head hurting you like when you lose the stud of your shirt in the middle of the bed.

— Flann O'Brien, *The Dalkey Archive*

7.1 Introduction

The significance of ...

Detailed tables

Xyz

Program code

Xyz etc

Glossary

Entries are listed in alphabetical order.

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.

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.

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.

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.

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.

List of Notation

Entries are listed in the order of appearance. The "Ref" is the number of the section, definition, etc., in which the notation is explained.

Symbol	Description	Ref
$\mathbb{F}_q$	Finite field of q elements	1.2

Index