Improved Peer-to-Peer Lending Credit Scoring Mechanism using Machine Learning Techniques

Tshauambea Murendeni

Student No: 11612709

Supervisor: Dr S. Moyo
Co-supervisors: Mr N. Mphephu

This research proposal is presented as partial fulfiment of the requirements of the degree:

Master of Science in Applied Mathematics
at the University of Venda
School of Mathematical and Natural Science

18 June 2021

Declaration

*I, **Tshauambea Murendeni**, declare that this research proposal titled: **Improved Peer-to-Peer Lending Credit Scoring Mechanism using Machine Learning Techniques**, submitted in partial fulfilment of the requirements for the conferral of the degree **Master of Science in Applied Mathematics**, from the University of Venda, is my own work and that all the sources that I have used or quoted have been indicated and acknowledged by means of complete references.*

I further declare that I have not previously submitted this work, or part of it, for examination at University of Venda for another qualification or at any other higher education institution.

18/06/2021

Mr M. Tshauambea *Date*

(Student)

18/06/2021

Dr S. Moyo *Date*

(Supervisor)

18/06/2021

Mr N. Mphephu *Date*

(Co-Supervisor)

Abstract

Peer-to-Peer(P2P) financing is a fast developing modern financial exchange network, which bypasses conventional intermediaries by linking lenders and borrowers directly. However, the online P2P lending platforms are faced with a problem of information asymmetry between lenders and borrowers. Assessing borrower's creditworthiness is important because many P2P loans are not secured by collateral. Banks use credit scoring to evaluate borrower's creditworthiness and reduce potential loan default risk. However, in P2P lending platform effective credit scoring models are hard to build due to insufficient credit information. This work is based on an empirical study by using the public dataset from the LendingClub, one of the largest online P2P lending platform in the USA. The aim of this study is to investigate the influential factors on loan performance on the basis of the credit score in the online P2P lending industry. This work improves the online credit scoring models and gives insight into the specific determinants that are influential for the score.

Keywords: *Machine learning, P2P lending, credit, creditworthiness, credit risk, credit scoring and information asymmetry.*

Acknowledgments

Ahead of everything, I would like to thank the almighty God for giving me the life, knowledge and strength to carrying out this research. I would also like to gratefully acknowledge my supervisors Dr Simiso Moyo and Mr Ndivhuwo Mphephu for their untiring efforts to guide me when carrying out the work presented here. Finally, I would like to thank the DSI-NICIS National e-science Postgraduate Teaching and Training Platform (NEPTTP) for selecting me to be part of this programme.

Contents

Chapter 1

Introduction

Banks assess the potential credit riskiness of borrowers by applying a method of credit scoring that uses scores to approve loans. The banks use statistical models to evaluate the credit intended for Small and Medium Enterprises(SMEs) in an automated, consistent and objective manner. Recently the use of statistical models has increased significantly in banks. However, some banks have continued to use the experts judgment in approving loans wherein the success of the approach relies on both the borrower and the bank's discretion (Board, 2017).

There are two stages where banks use credit scoring model to determine credit risk, namely: the application selection process by introducing a minimum cut-off score based on the level of risk and return on investment, and the performance measure of the SMEs once their loans have been secured by calculating the behavioural scoring. Banks use credit scoring for the following significant advantages: physical interaction between banks and customers is limited, and speedy approval of loans based on borrower's electronically submitted documents and credit history (Siddiqi, 2017).

Online Peer-to-Peer(P2P) lending involves offering of services of matching lenders and borrowers, which is a well known practice of lending money to individuals. The number of P2P lending operators and loan volumes are steadily growing fast globally. In P2P lending lenders stand to earn higher returns for their investments, than they could make from banking institutions. However, there is always higher risk of borrowers defaulting on loan repayment (Lynn et al.).

The process in P2P lending is that the borrowers apply for funding at the lending platform. The platform does the credit risk analysis and then sends their findings to the lenders without verifiable information about the potential clients. The lenders select the loan applications

they would like to invest in and put a competitive bid it in order to get the funding opportunity. The provision of funds and repayments are made via the client's account created by the platform for the purpose of interaction between the borrowers and lenders (Board, 2017). Figure 1.1 demonstrates the P2P lending process and the interaction with each participant.

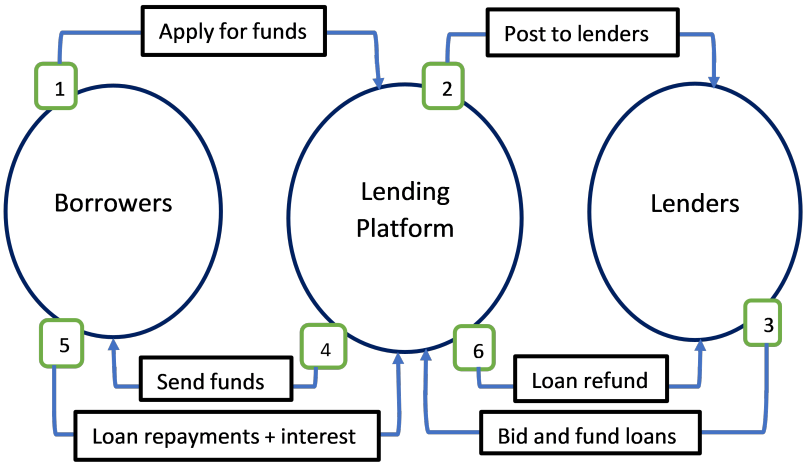


Figure 1.1: Connection of Lenders and Borrowers through FinTech platform

The P2P lending platform is similar to other platforms market based which allow for buyer and seller interaction, for example Uber transportation. However, the P2P lending platform differs from these other market based platforms because the operators in this case provide quality assessment of the loan application and manage the obligation of the borrower to the lender as well as provide the lenders with an account management service (Davis and Murphy, 2016). Similar to loan application in traditional banks, borrowers apply for funding for different purposes in the P2P lending platform. The interest charged on the borrower's loan is divided into two fees namely, the platform transaction usage fee and investment profit of lenders (Board, 2017).

The P2P lending platforms calculate their own credit ratings as opposed to banks. The banks use professional credit institutes for that purpose. However, credit grades from P2P platforms are less accurate as opposed to credit gradings from the traditional credit institutions. In P2P platform interest rate for individual loans is charged on the basis of their credit grading (Möllenkamp, 2017). For the LendingClub lending platform, interest rates range from 6% to 33%. This study analyses the loans that already matured, and as a result there are only two possible outcomes: the funded amount plus interest charged were fully paid back or the borrower defaulted on their loan repayments by stopping their payments for more than 121 days.

Currently there are over 47 P2P lending companies on the weblog P2P-Banking.com existing worldwide. Zopa was the first Europian lending platform established in 2005. In the USA the P2P lending industry started with the launch of Prosper in 2006, then followed by LendingClub. Smava was the first P2P lending company in Germany established in 2007. Wanga.com, Rainfin and Lendico are currently the South African social lending operators (Bachmann et al., 2011).

The online P2P lending has several advantages over the traditional banks. These include the less than 14 days taken for the funding to be concluded, funding process is straightforward, borrower can simply access funding on the internet and lastly lower interest rates on the loans (Möllenkamp, 2017). The online P2P or crowd lending is gaining acceptance through out the world because of its unique ability to make lending to poor people and non-transparent small and medium-size enterprises.

The online P2P lending has some drawbacks faced by the lenders, and the default risk of the borrower is the most common one. Online P2P lending companies transfer the credit risk to lenders who in turn grant the loans at their own risk. The online P2P lending is more risky than traditional banks lending. Online P2P lending loans are unsecured, because there is no collateral involved. In most literature default risk and loan performance are closely associated with the differences in the quality and quantity of available information between borrowers and lenders called information asymmetry (Lynn et al.).

This study focuses on the decision problem faced by the online P2P lending platform called the LendingClub in deciding whether to accept the loan application or not. For the LendingClub to make such decisions they should be able to distinguish between a bad and good payer, which is a dichotic response variable but could be easily transformed to a binary classification problem. Each loan in the LendingClub is rated on a scale of seven grades from A to G, which respectively increase with the increase in default risk.

The statistical approach of credit scoring summarises the available information to a score, which is used to determine the credit worthiness of an applicant. A lot of statistical credit scoring techniques have been proposed in the literature, but none of them have been proven to be the best for all forms of data. Only matured loans are considered in this study and therefore the loans performance can only be 'default' or 'fully paid'. Since just two outcomes are possible, the logistic regression is dichotomous. 'Default' is set as $y = 0$, 'fully paid' is set as $y = 1$. The binary credit scoring models will be used to accommodate the categorical dependent variable.

This study argues that some independent variables negatively affect the performance of the

credit scoring model, while others improve it. These variables determine the score which is used to differentiate between bad payers and good ones. The assumption made in this study is that the independent variables that influence the loan performance will also alter the credit scoring model's ability to accurately predict the score.

The number of independent variables that are provided in the loan statistics of LendingClub is quite large. However, not all are of interest for this study. The variables of interest for this study are selected on two conditions namely: the predictive powers of each variable given by the information value(IV) which range between $0.02 \leq IV \leq 0.5$ and the improved credit scoring.

In order to determine the loan performance of the improved P2P loans of different credit grades through a credit scoring model, the significant explanatory variables that minimise the number of defaulting loan applicants must be determined. For every set of explanatory variables the binary logistic regression will be run to calculate the accuracy of the predictive model in reducing the number of defaulting borrowers.

Chapter 2 deals with literature reviewed in relation to the hypothetical questions raised in connection with the topic of the study. In chapter 3 the problem of this research is described and the key objectives are stated. This is followed by the chapter dealing with the data cleaning in order to allow for the use of that data in the context of our problem and problem solving techniques used. Chapter 5 describes the methods used in achieving the improved models for attaining improved credit scoring mechanism. Lastly, the results of the implementation of the above specified methods are analysed and discussed, and conclusions are drawn with clear indication of future studies.

This study used the Python programming language. Python is an open source and free to use programming language. It has Pandas library with strong data analysis tools better suited for use in repeated tasks and data manipulation.

Chapter 2

Literature review

2.1 P2P Lending

P2P or crowd lending is gaining acceptance world wide because of its unique ability to make lending to poor people and non-transparent small and medium-size enterprises. It is characterised by what many banking institutions would consider high-risk, such as small and medium-size enterprises loans. Individuals with low revenues relie on P2P lending for external financing. In the P2P lending there is a borrower, lender and the lending platform. The borrower could be individuals or small businesses looking for reasonable rates for refinancing debt, but struggling to get a loan from the conventional banks. The lenders are individuals looking for higher rate of return on investment. In simple terms, they are profit-driven (Xing and Marwala, 2018).

The lending platforms are different in their basic lending models. In some platforms, the loans are selected by the lenders based on the loan purpose, borrower's income, loan term and other credit indicators. Whereas in some platforms the credit risk assessment is done from within the platform and also contributes to the loan selection. The lending platforms use more data sources that are non-traditional when compared to banks , for example, the use of online spending behaviour by the Indian P2P platform (Board, 2017).

The lender's risk involved in the P2P lending is minimized by the level of credit-risk of the borrower. It is very important to know if the borrower is credit-worthy or not. The most likely cause of credit risk in P2P is information asymmetry that exist between the borrower and the lender. Lenders tend to "learn by doing", since credit-risk are sometimes hidden and too complex to overcome (Xing and Marwala, 2018). The lending platform encourages the lenders to invest in multiple loans in order to spread the risk associated with information

asymmetry. On the other hand, online P2P deals with credit risk by associating themselves with credit agencies to gain more credit information on the borrower, assign a credit risk level to each approved loan request, limit the amount a borrower could get based on their credit risk level and hire a collection agency to represent the lender in case of a borrower failing to repay the loan.

Currently, P2P lending platforms provide a service of managing investments for individuals, which allows for earning high returns with a streamline process by lenders, while borrowers enjoy fast approvals of their loan applications with low interest rates.The P2P lending uses two operation models in order to be more accommodative. The models are auction-based lending and automatic-matched lending (Milne and Parboteeah, 2016).

For lenders to decide which loan proposal is worth investing in, they must consider various issues such as the trading efficiency, loan performance and risk. Researchers provide relevant recommendations by developing models based on two roles called assess loan requests and investment offers based on multiple objectives, portfolio selection and optimization (Fong, 2015).

An operator's main responsibilities are assessing credit, loans administering, providing the gateway for requesting and bidding process, providing legal assurance, and coordinating the payment process for successful loans (Larrimore et al., 2011).

In general, the P2P lending platforms have several advantages over the banks. The lenders and borrowers conduct business directly leading to better investment opportunities resulting in lower costs of conducting business for borrowers. Borrowers who normally would not get credits from banks have an opportunity of securing a loan through these platforms. The platforms are more responsive and pay a lot of attention to social values than in the traditional banking system. They provide improved technological supported quality service and speeded processes in respect to lenders and borrowers. They also provide additional services such as recommendation of loan match, early repayment options and quick funding, and lastly the protection of borrowers against any embarrassment (Milne and Parboteeah, 2016).

P2P lending platforms consist of two diverse groups of of activities which include acting as a middleman and as a match-maker. These activities are flexible making it possible for the P2P operators to implement different lending model options. The lending models can be divided into the following categories: Auction-based and Automatic-matched lending. The LendingClub, Prosper and Zopa P2P lending platforms implement the auction-based lending. The Kabbage P2P lending Platform uses the automatic-match lending (Milne and

Parboteeah, 2016).

In P2P lending, the borrowers and lenders set the range of interest rates they are comfortable with on the loan. The platform then follows through by making the lenders bid in an auction for the loan within that range of interest rates the borrowers are willing to pay. The platform may also set the interest rates based on credit risk grade assigned to the loan, but with incentives to change it in case, the borrowers and lenders are not in agreement. Lastly, the borrowers could obtain the indicative rate from the online market based assessment of their risk profile, and lenders would then compare loan options within the platform (Davis and Murphy, 2016).

Some lending platforms monitor the use of the loan after the provision of the funds to avoid additional risk of clients missing repayments. For sufficiently late repayments the platform engages the debt collectors to recover the loan, which result in higher fees for the investor at the end of the collection. In South Africa, some platforms provide a secondary market that allows lenders to withdraw their offer of funding for some fee, provided other lenders are interested in buying the underlying loans (Board, 2017).

In P2P lending the major issue is information asymmetry. This is a major disadvantage in making a loan decision since the lenders only know what the platform chooses to reveal about the borrowers. In some cases this is so because the information which will have been provided by borrowers would have been incomplete, whereas the platform could also decide to keep some information confidential for other reasons. It must also be noted that the credit grades allocated to loan applications are not a guarantee that the associated risk is low, but rather this is done on the basis of the fact that the probability of not defaulting is sufficiently high (Lynn et al.). This is the reason why most platforms advice the lenders to invest in more than one loan, to spread their risk (Board, 2017).

The grade assigned by the P2P lending platform together with the borrower's debt information plays an important role in predicting default. For more strict P2P lending platforms, the lower the grade assigned, the higher the interest rate charged on the loan because of the associated higher risk of defaulting involved. The P2P lending platforms rely on the FICO score from the third party to assign their grades to the loans. The lendingClub platform affirms that they use a formula that uses the FICO score and some information provided by the borrower. However, the formula and the borrower's information to compute the P2P lending grades is kept secret from the public use (Serrano-Cinca et al., 2015).

Emekter et al. (2015) conducted an empirical research on P2P LendingClub platform dataset and concluded that the FICO score and debt information improves the accuracy of loan

default prediction model. In recent years, P2P lending platforms have become the data sources of credit scoring (Niu et al., 2019). Bachmann et al. (2011) pointed out that the difference in online P2P lending platforms lies with the lender's expectation for returns, which should determine whether the platform is commercially viable or not.

In 2011 Bachmann et al. (2011) conducted literature review study on the online P2P lending and listed 10 biggest lending companies based on the volume of loans created; the LendingClub was on the fifth position. Later on in 2015, Serrano-Cinca et al. (2015) identified the LendingClub as the largest US P2P lending company, in their empirical study of finding the determinants of default in P2P lending.

2.2 Credit scoring

The word credit itself means "buy now, pay later" (Anderson, 2007). If the borrowers are willing and showing the ability to pay, they are said to be creditworthy. A change in the borrower's creditworthiness caused by any financial irregularity is called the credit-risk. Giving credit involves trusting the client or customer, however in lending the lenders increase the charges to cover the risk if the trust is low. Ranking of items according to their quality in order to differentiate between them and making consistently objective decision thereafter using numerical tools is referred to as scoring. Predictive scoring models use past experience to predict the relative most likely future events.The outcome expected from scoring is a score indicating the level of risk.

The credit scoring formulas are different and, accordingly, produce different ranges of scores, for example the FICO score ranges from 300 to 850, the PLUS score ranges from 330 to 830 and the Vantage score ranges from 501 to 990. The similarity of the scores produced by all these formulas is that the higher score represent a customer with sufficiently high probability to pay back a debt. The lower the credit score the higher the interest charged on the loan, so it is important to have a higher credit score so that the interest accompanying the loan is lower (Mathew, 2017).

Credit scoring is a way to evaluate credit-risk based on the borrower's or customer's historical payments and make consistent credit decision about the borrower. The benefit of credit scoring is that the time taken for both analysing creditworthiness and better decision-making process is minimized. Currently the credit scoring has been transformed into binary or dichotic classification and the credit scores have been used to decide whether the loans could be granted or not. The higher the credit score the borrower obtains, the higher the chance that he or she will get the loan. The key assumption in building a credit scoring

model is that the future resembles the past, meaning that every credit scoring depends on the historical information (Xing and Marwala, 2018).

The methods for credit scoring can be divided into two categories namely: the technological and behavioural/judgemental credit scoring methods (Xing and Marwala, 2018). The technological credit scoring methods are based on statistical and mathematical techniques to evaluate the credit risk, for example, logistic regression, etc. The behavioural credit scoring methods are based on the initiative by the borrower to subject oneself to evaluation of their credit risk by a credit expert. The benefit of involving behaviour in credit scoring is to undercut lying and limit the uncertainness of the lender associated with borrower's behaviour.

The behavioural credit scoring does not use historical information on the borrower only just like in the case of the technological scoring. This credit scoring uses the borrower's character and integrity, the difference in the borrower's assets and liabilities, the collateral provided in case payment problems occur the borrower's ability to pay, and the borrower's circumstances. Other literature refer to these borrower's information as the five C's, standing for characters, capital, collateral, capacity, and condition (Baesens et al., 2016; Anderson, 2007). Behavioural credit scoring is not very reliable since it depends on the experience of an expert. The expert in the field is a human, and this leads to longer time required to produce results which in some cases are less accurate and inconsistent.

The shift caused by credit scoring from relationship lending to transactional lending left small lenders holding on because they believe relationship gives them competitive advantage. Again credit scoring caused the shift from collateral and guarantees in secured lending to information and future events in unsecured lending. There are cases where the scorecards are insufficient because the potential risk is high and the customer is fighting against the system decision. In such cases, a credit expert judgment tends to be the best option to provide transparent assessment so the borrower can understand what transpired in coming up with the decision to reject the loan. Today credit scoring is used in markets including unsecured, secured, store credit, service provision, enterprise lending, etc (Anderson, 2007).

The unsecured market includes personal loans, credit cards, and overdrafts, whereas for secured market we have home loans mortgages and motor vehicle finance. The store credit market consists of furniture, clothing and mail order. The service provision has to do with municipal accounts, short-term insurance, and phone contracts. Working-capital loans and trade credit are included in the enterprise lending market. Credit scoring improved the unsecured and store credit market the most since they both heavily depend on information (Anderson, 2007).

The source of credit scoring information in the data collected from the borrower is loan features, borrower's personal information, optional information, and social information. The loan features are loan title, purpose, amount, duration, quality, period, and so on. The borrower's personal information is a bank account, assets, credit grades, income tax returns, and debt to income ratio. The optional information is the borrower's picture, however, it could be used to influence the lender's decision and lead to discrimination. The social information is repetitive interactions in social networks, to supplement the aggregated financial information of the borrower (Xing and Marwala, 2018).

In building credit scoring model there is no fixed number of variables to be used since they differ from one study to another and depends on the data provided. In some studies, only twenty to thirty variables are used but some use more variables, based on the nature of the data. The models themselves have no fixed score point to cut-off good customers from bad ones, so that determination lies with the analyst. The above assertion implies that the analyst will determine how much they are willing to risk, based on the cut-off score point he/she set (Abdou and Pointon, 2011).

The statistical techniques used to build an effective and efficient credit scoring model are logistic regression, decision trees, discriminant analysis, probit analysis, neural networks, support vector machines, regression analysis, the weight of evidence measure, linear programming, geneticalgorithms, Cox's proportional hazard model, k-nearest-neighbour and genetic programming. In all of the above mentioned statistical techniques, there is no best technique for all data sets. The advanced statistical techniques have higher predictive ability than the conventional statistical techniques, but both have similar predictive capabilities (Abdou and Pointon, 2011).

Henley and Hand (1996) compared the results of the KNN with those of the logistic regression and decision trees, the results were relatively close. The logistic regression technique can fit different distribution functions and fit the credit scoring problems well (Abdou and Pointon, 2009; Mpofu and Mukosera, 2014; Abdou et al., 2008). West (2000) compared five neural network models against the traditional methods, the logistic regression has the lowest credit scoring error for both Australian and German credit data.

In a lot of literature presented in Paliwal and Kumar (2009), the confusion matrix was used as the performance evaluation criteria. This performance criteria play an important role in credit scoring since it highlights the accuracy of the model's predictions (Abdou and Pointon, 2011). The Receiver Operating Characteristics (ROC) performance criteria minimize the total misclassification costs, by giving the cut-off point (Abdou and Pointon, 2011).

Chapter 3

Problem setting

3.1 Introduction

This study is based on analysis of data obtained from P2P lending investment platforms. A data scientist must provide sound advice to potential investors who are looking for opportunities to invest their money through funding loans in the platform. This involves being able to calculate the riskness of the business. Over the years, the platform has managed to build a solid reputation with lenders and borrowers. To protect its reputation and retain clients, it is important that the platform's risk assessment mechanism of the potential borrowers is robust and effective so that the default rates are kept low. It is important to remember that the loans that are given out are unsecured making it apparent that bad loans should be avoided at all cost. Earlier, the platform calculated the credit scorecards based on regression techniques and now those scorecards have been shown to be underperforming. Against this backdrop, the platform deemed it fit to build new efficient scorecards to assess new clients.

Over the years of operation, the platform has built a database profiling all its customers. Now, this information needs to be used to improve a credit scorecard. Given the aggressiveness of machine learning and artificial intelligence techniques in dealing with large data sets the platform's administrator sought to utilize these methods in improving on scoring mechanism at the platform. The primary objective of using these new techniques will be to develop an efficient scoring mechanism based on the collected business activities data at P2P lending platforms.

3.2 Problem statement

Lenders invest in loans in order to earn profit. However, due to information asymmetry, where one party has better information about the transaction than the other party they are dealing with, some investments end up being a loss. In banks loan applicants are monitored with the help of credit institutes in order to secure loan repayment. In P2P lending, the asymmetric information flow from borrowers to the lenders makes monitoring loan applicants and judging the credit risk of the loan difficult, when compared to the traditional banking systems. To determine which loans are profitable investors or lenders must know which variables must be closely monitored from time to time.

This study looks at predictor variables associated with a credit scoring model with a view to assess how these factors affect the profitability of a loan in a P2P lending process through optimisation of the scoring function. According to Möllenkamp (2017) the most influential independent variables for the P2P lending platform loan performance are the platform assigned loan grades, the loan amount, the borrower's annual income, debt-to-income ratio and inquires in the last 6 months.

3.2.1 Aim of the study

The main aim of this study is to minimise the chances of the loan applicants defaulting on repayments through improved credit scoring mechanisms leading to improved prediction models predicated on machine learning approaches.

3.2.2 Research Questions

- We shall seek to improve the credit scoring models for Peer-to Peer lending using machine learning techniques so that interested investors are able to identify good borrowers from bad ones.
- We shall seek to establish whether there is an optimal number of predictor variables that must be used in a credit scoring model.
- We will also work to establish whether filtering variables could improve the current score by reducing the risk of borrowers defaulting on loan repayment.

3.2.3 Objectives of the study

- To filter the variables in order to remove the variables that negatively affect the performance of the credit scoring model.
- To use algorithms based on improved credit scoring models to improve on the accuracy of the calculations of the predictive variable.
- To optimise credit scoring function in order to minimize the number of loan defaulters.
- To compare the standard credit scoring method with the improved credit scoring methods.

3.3 Rationale for carrying out the research

It is very important for the investors or lenders to know which variables are relevant for determining which loans are worth investing in. There has been a number of studies conducted on the LendingClub data sets, and most of them investigated the determinants of the P2P loan performance. However, these studies on P2P lending did not analyse the loan performance on the basis of the credit score. This work focuses on improving the credit scoring model and gives insight into the specific determinants that are influential for the score.

This work is the first of its kind which looks at influences of different variables on the score to improve the loan performance. This study assumes that an effective credit scoring model will in turn lead to improved loan performance. Five studies done by Carmichael (2014); Serrano-Cinca et al. (2015); Emekter et al. (2015); Li et al. (2016) and Möllenkamp (2017) focused on the LendingClub data and analysed the determinants of loan default with similar approaches and found different results.

The difference in the results of the above studies, is in the number of variables listed as most influential loan default determinants. The discrepancy in results can be alluded to the fact that the authors used different variables and methods in their studies. The other reason, though it might not carry much weight, is that these studies mentioned are using data in different years and the loan default determinant throughout the years might be different.

Carmichael (2014) and Emekter et al. (2015) used 36 and 60 months loan term on the LendingClub data set and also included the FICO score as one of the determinant of loan default. In our study the last FICO score will not be used since it depends on the loan outcome and using it in a predictive model would result in overly optimistic results.

All of the five studies mentioned above used the LendingClub data set recorded over a period

of less than 10 years. This study is the first one to use the data set recorded over a period of 11 years. In this study we expect that the results might be slightly different to those of the afore mentioned studies because of situations due to unusual economic conditions like the financial crisis during the 11-years period data recording.

Chapter 4

Exploratory Data Analysis

The data set used in this study is publicly available at www.lendingclub.com. The data set is from 2007 to 2018, and has two files namely accepted and rejected loans. The accepted file has 2260701 observations and 151 variables.

4.1 Data cleaning

4.1.1 Feature selection

To clean the data set the weight of evidence(WOE) and information value(IV) will be used to explore the quality of data and screen variables. In the credit scoring world the WOE and IV have been used to explore data and screen variables. The WOE is a measure of attributes to differentiate between a portion of good and bad accounts. The formula for WOE is given below by taking the natural logarithmic of the ratio of the distribution of good to bad.

$$\text{WOE} = \ln\left(\frac{\text{Distribution of good}}{\text{Distribution of bad}}\right).$$

In terms of a good account being named as an event and bad account as a non-event, the WOE formula assumes the form

$$\text{WOE} = \ln\left(\frac{\%\text{of non-events}}{\%\text{of events}}\right).$$

The IV is calculated by the formula given below as

$$\text{IV} = \sum(\%\text{of non-events} - \%\text{of events}) \star \text{WOE}.$$

According to Siddiqi (2017) the information values could be used to filter variables in accordance with the classification in the table below:

Information value(IV)	Predictive power
IV $<$ 0.02	too low
0.02 $\leq$ IV $<$ 0.1	Weak
0.1 $\leq$ IV $<$ 0.3	Moderate
0.3 $\leq$ IV $\leq$ 0.5	Strong
IV $>$ 0.5	Too Strong

Table 4.1: The IV range and their predictive ability.

The LendingClub dataset have 151 variables and most of these variables are updated as time goes by. This means that most variables were not available at the time the loans were issued. All the variables that are updated with time will be removed to avoid model leakage. Model leakage occurs in a situation where the model is being applied with data that would not have been available at the time when the model was used to predict a phenomenon. Model leakage can happen if some or just one variable and the target variable are highly correlated.

This study used table 4.1 to filter independent variables from 152 to 62. The variables for which $IV > 0.5$ were removed, because they are associated with model leakage and correlate highly with the target variable. The variables with their IV less than $IV = 0.01$ were also removed, because they have no predictive ability or they have no effect on the dependent variable predictions. Figure 4.1 below shows all the viariables that were removed because their IV were grater than 0.5.

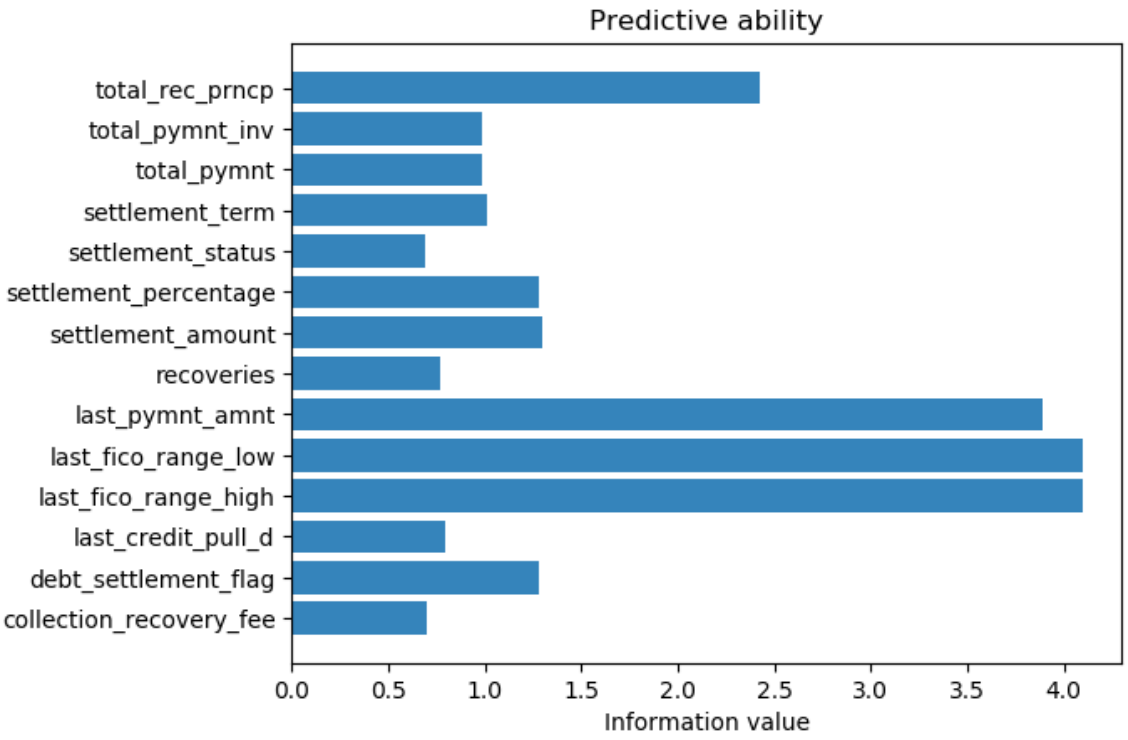


Figure 4.1: Model leakage associated variables.

4.1.2 Missing values

When working with real life data, there is always going to be missing values. This could be caused by the following but not limited to: information was not needed at that time or not available at the point of data collection. Later this information would be needed for recording. Since, most machine learning algorithms cannot handle missing values in the data set, the approach to deal with the missing values depends on their quantity. The missing values in categorical variables are mostly replaced with mode value, while those in numerical variables are replaced by mean or median value. However if the missing values are too many dropping those observations, to prevent contaminating the results is normally the recommended way of dealing with the situation.

The LendingClub dataset has 15 independent variables that had more than 90% of their entries missing. These independent variables are: “debt settlement flag date”, “deferral term”, “hardship amount”, “hardship dpd”, “hardship end date”, “hardship last payment amount”, “hardship length”, “hardship loan status”, “hardship payoff balance amount”, “hardship reason”, “hardship start date”, “hardship status”, “hardship type”, “payment plan start date”, “settlementdate”. These independent variables were all removed from the dataset, because there was a lot of missing information.

Some independent variables namely: “all util”, “inq last 12m”, “open acc 6m”, “open rv 12m”, “open rv 24m” had 38.32% missing entries and very weak predictive ability with an IV between 0.021 and 0.040. Dropping variables with 38.32% missing entries contaminates the whole cleaned dataset. However, we take note of the fact that removing these independent variables does not affect the information on the dataset. The variable “mths since recent inq” is independent of the method used to handle its 13.068% missing data entries, since the accuracy of the standard scoring method remains the same.

These processes reduced the number of independent variables much further from 62 to 32 variables. All the loan statuses that had not reached marturity stage or expired at the time of data collection were removed. This resulted in 915318 rows being removed leaving a total of 1345350 rows of martured loans. Throughout the data cleaning process the most critical decision was to make sure the predicting ability of the standard scoring method is not compromised. This was made possible by making sure that before removing a variable, the information it carries is small enough not to affect the accuracy of the model.

The independent variables left after the data cleaning processes are: “acc open past 24mths”, “annual inc”, “avg cur bal”,“bc open to buy”, “bc util”, “dti”,“funded amnt”, “funded amnt inv”, “grade”, “home ownership”, “installment”, “int rate”, “loan amnt”, “mo sin old rev tl op”, “mo sin rcnt rev tl op”, “mo sin rcnt tl”, “mort acc”, “mths since recent bc”, “num actv rev tl”, “num rev tl bal gt 0”, “num tl op past 12m”, “percent bc gt 75”, “revol util”, “term”, “tot cur bal”, “tot hi cred lim”, “total bc limit”, “total rec int”,“total rec late fee”, “total rev hi lim”, “verification status”, and “loan status”

4.2 Data representation

The LendingClub lending platform offers loans to borrowers for different purposes. The diagram below shows a cloud of loan applicants purposes for which the loans were applied. This cloud of words measures the frequencies of how many times a word has been used. The word font size increase corresponds to the frequency it has been used.

Figure 4.2: The cloud of loan purposes

The words in figure 4.2 are in different font sizes, the ones appearing in big font size represent frequently used purpose for a loan and the ones in small font are less likely reasons for a loan. To get a clear picture about which loan purposes influence the default rate, we pick 12 most dominant loan applicant purposes and draw a comparison histogram diagram as shown below.

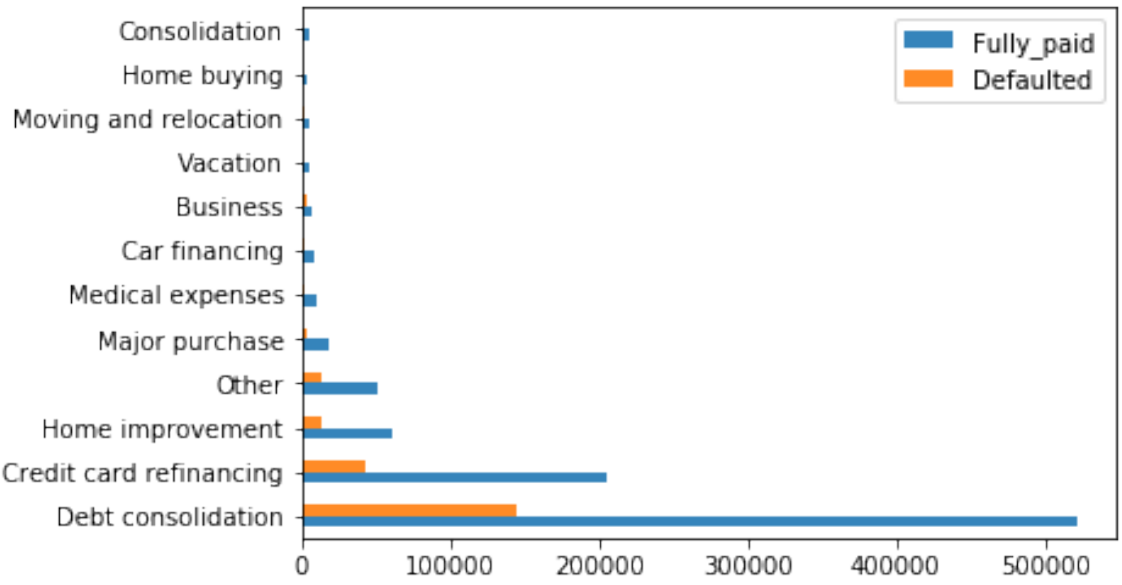


Figure 4.3: Loan performances based on loan purposes

In figure 4.3 the most dominant loan purpose that a number of loan applicants used in their loan application is "Debt consolidation". There is roughly 53% of the loans that were issued by the LendingClub with the purpose of consolidating debt, and only 22% of those loans resulted in default. The table 4.2 below illustrates the percentages of borrowers who are fully paid up and the defaulters against the percentage of purpose. Contribution of each group calculated out of 12 most dominant loan purpose contribution groups, is given by the percentage of the ratio of number of loan purpose per group to the whole cleaned dataset.

Loan Purposes	Fully Paid	Defaulted	Purpose total contribution
Debt Consolidation	78.33%	21.67%	52.78%
Credit Card refinancing	82.53%	17.47%	19.73%
Home improvement	81.67%	18.33%	5.90%
Others	78.42%	21.58%	5.13%
Major purchase	79.97%	20.03%	1.86%
Medical expenses	77.52%	22.48%	1.05%
Business	69.17%	30.83%	0.90%
Car refinancing	84.19%	15.81%	0.87%
Vacation	80.46%	19.54%	0.61%
Moving and relocation	75.45%	24.55%	0.61%
Consolidation	85.40%	14.60%	0.47%
Home buying	76.88%	23.12%	0.45%

Table 4.2: The summary of loan purposes contributions.

The percentage of the loan purpose contribution out of the total purposes loans contributions declines depending on the type of loan purpose. The highest percentage is recorded for debt consolidation followed closely by credit card refinancing and home improvement. Home buying contributes the lowest percentage followed closely by consolidation. In the middle we have business followed by medical expenses. The default rate shows no relationship with the loan purposes total contribution. Table 4.3 below shows the LendingClub platform loan applicant's overall perfomance statistics measured by whether the loan was fully paid or not at the end of term of the loan.

Martured_Loan_status	Total_count	Total_percentages
Charged Off or Defaulted	255235	20.201%
Fully Paid	1008239	79.799%

Table 4.3: The summary of loan applicants perfomance.

Based on our assumptions and simplifications for this study, table 4.3 shows that 80% of the loans issued from 2007 - 2018 by the LendingClub were fully paid back. The LendingClub offers loans for only two loan repayments periods, the 36 months and 60 months. It is therefore interesting to understand the loan performances in those terms. The table below summarizes the performances per term.

Loan_term or Period	Overall Number of Applicants (Percentage)	Percentage of Term Defaulted Loans	Percentage of Term Fully Paid Loans
36 months	75.796%	16.168%	83.833%
60 months	24.204%	32.833%	67.167%

Table 4.4: The summary of loan terms perfomance.

Table 4.4 shows that about 76% of the loans issued by the lendingClub platform from $2007 - 2018$ were of 36 month repayment period. Only about 16% of these 36 months loans ended up in default. From table 4.4 it is clear that not many lenders invested in longer period active loans, which might be the reason why the 60 months loans contributed just 24% of the loans issued. The LendingClubs classify the loans according to grades, A, B, C, D, E, F and G, based on associated risks. The following diagram provides a visual aid for the number of loans issued for 60 months and 36 months terms in each grade assigned by the LendingClub platform.

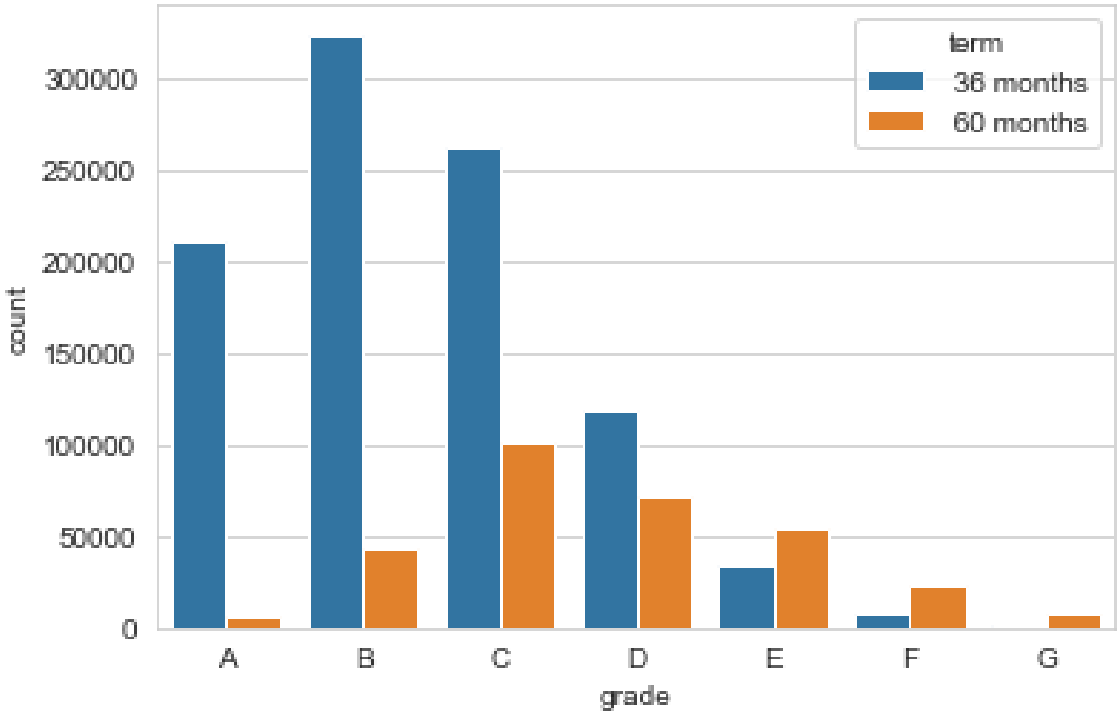


Figure 4.4: Number of loans issued in different terms per grade.

Figure 4.4 shows that the 36 months loans are dominant in grades A-D. For grades with less default risk, investors are funding 36 months the most. However, it appears as if as the default risk rises the investors allow for more repayment time. It looks more like investors are using time to minimize the chance of default risk. In other words, the investors notion is that the longer the period given to the borrowers to make repayments, the less likely it would be for them to default. Nevertheless, the choice of a loan term lies with the borrower, based on the two options of loan terms provided by the lending platform.

The lending platform assigns grades based on the default risk estimated from the information supplied by the borrower in the loan application process. That default risk increases from grade A to grade G. To confirm the behaviour of the default risk, we construct a table 4.5 showing default risk of the loans issued per grade.

Loan grade	Loan Status	Total Counts	Grade percentage
A	Fully Paid	204053	93.995%
···	Charged off or Default	13037	6.005%
B	Fully Paid	318537	86.571%
···	Charged Off or Default	49410	13.429%
C	Fully Paid	281188	77.3683%
···	Charged Off or Default	82253	22.632%
D	Fully Paid	130827	69.191%
···	Charged Off or Default	58255	30.809%
E	Fully Paid	53352	60.860%
···	Charged Off or Default	34311	39.140%
F	Fully Paid	16114	54.119%
···	Charged Off or Default	13661	45.881%
G	Fully Paid	4168	49.174%
···	Charged Off or Default	4308	50.826%

Table 4.5: Loan applicants performance in lending platform grades.

There is no grade without default risk, however some grades have higher default risk. For example in grade G there is more than 50% chance of a loan to be charged off or default. From table 4.5 the default risk of a loan applicant seems to increase with grade assigned by the lending platform from A - G. The number of loans that are funded seems to decrease for grades D - G. However, this comparison of default risk between different credit grades does not gives us any useful information about trends on the number of applicants between grades since all grades recieved different number of funded loans and the default chance is calculated based on that number.

The conclusion we make from this is that making investment decisions based on table 4.5 would be unadvisable. The visual aid to support our claim in this study that grades received different number of funded loans and make grades perfomance comparison redundant is provided in figure 4.5 below. In addition, table 4.5 reveals clear increasing default rate corresponding to grades C to G. This provides vital information required by investors to know and understand the risk associated with various grades.

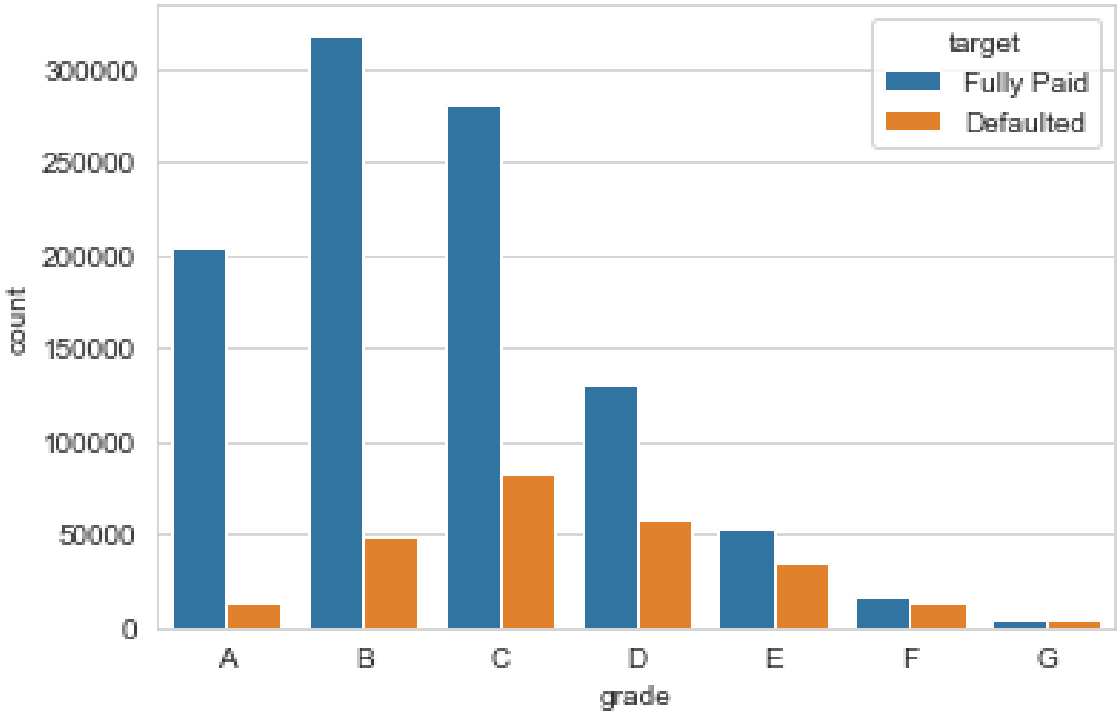


Figure 4.5: Loan applicants performance in each credit grade

The competition for funding loans is high amongst lenders funding loans in loan grades(A, B and C) which according to the lending platform have less risk of default. The borrowers apply for funding, and the platform assesses the default risk associated with each borrower based on the information supplied in the application for funding. On that basis a credit grade rating is allocated and the corresponding interest rate is charged.

The figure 4.5 is a bar chart data presentation depicting the number of applicants in each grade who fully paid their loans and those that defaulted. There are fewer applicants in grade A compared to grade B and C, presumably because of stringent conditions for qualifying to be in A than in the other grades. There are much more fewer borrowers in grades E to G, presumably, because there would normally be fewer investors who are prepared to take high risk by funding high risk borrowers.

The figure 4.6 below illustrates the five number summary per loan credit grade assigned by the platform based on the level of interests rates.

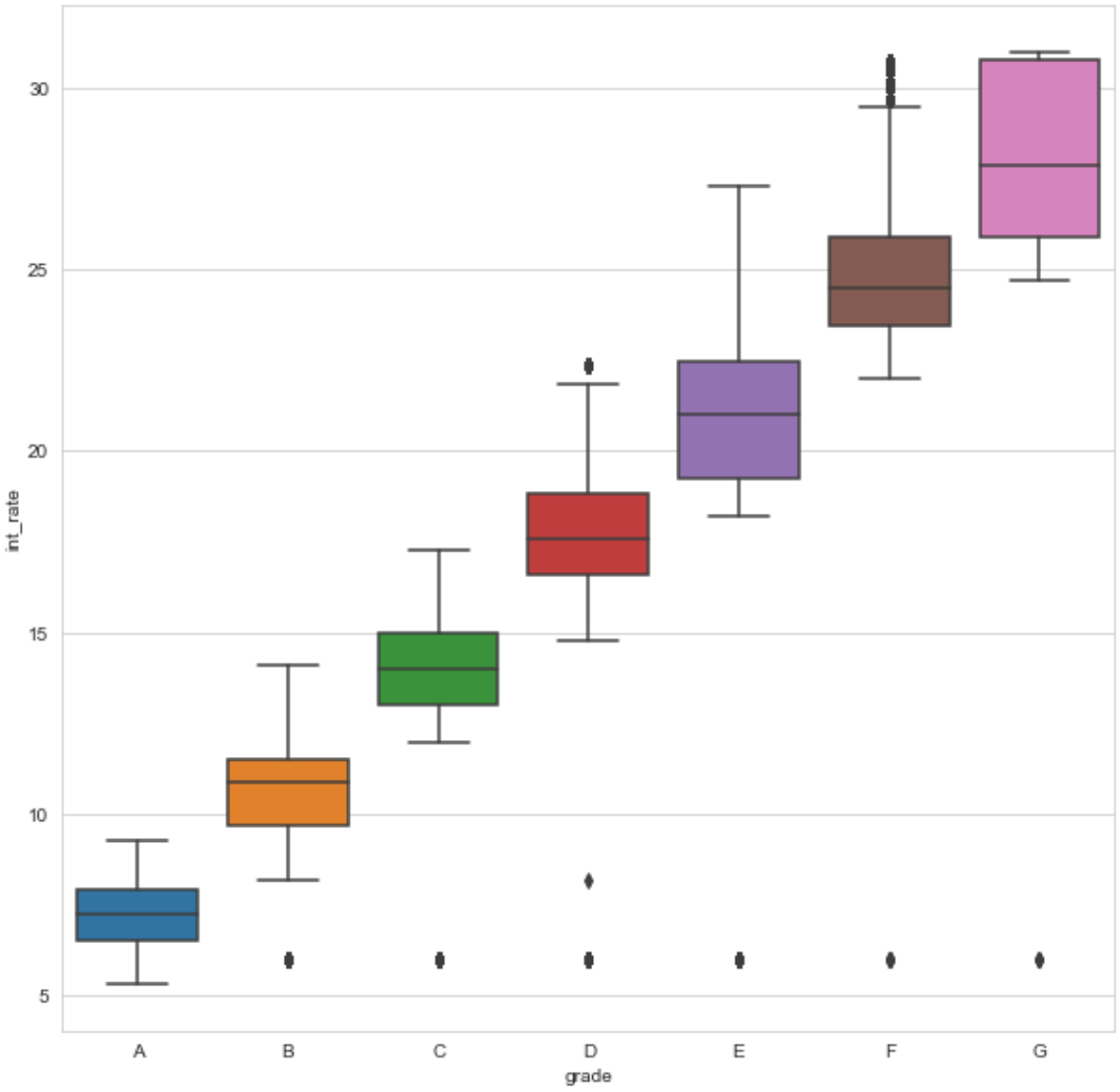


Figure 4.6: interest rates range in each loan credit grade

As expected the interest rates on the loans increase with change in credit grade ratings from A-G assigned by the lending platform. However, the higher the interest rate is the higher the return will be on the loan, but the borrowers in that category are less likely to repay the loan. The outliers in figure 4.6 are due to loans that had to be auctioned based on the fact that more than one lender or none of the lenders wanted to fund them. In other words, interest rates on the loans that are seperated from others are due to allowing all the competing lenders to contest for a loan or removing uninterested lenders from contesting the funding of a loan.

Chapter 5

Credit Score Allocation methods

5.1 Modelling framework

Credit worthness scoring involves distinguishing between bad borrowers and good borrowers, and is a supervised machine learning technique which could also be transformed to a binary classification method. In a P2P lending platform there are only two matured loan statuses, namely: the “default” and “fully paid”, making the dependent variable dichotomous and could be transformed to binary classification easily (respectively “0” or “1”). For the purpose of determining creditworthiness of individuals this study used several credit scoring techniques such as credit scoring model for individuals, logistic regression (LR), adaptive boosting(AdaBoost), random forest(RF), gradient boosted decision trees(GBTs) and the stacked ensemble method. This study compares the accuracy of the LR, RF, AdaBoost, GBTs and the stacked ensemble credit scoring models.

In a P2P lending platform, there are only two expected outcomes, namely: the “default” and “fully paid” loan statuses. Therefore, the dependent variable is dichotomous and is easily fitted to a binary model as shown below. Consider the data set given by $\boldsymbol{D} = \{\boldsymbol{X}_i^{(j)}, \boldsymbol{Y}^{(j)}\}$, where

$$\boldsymbol{X}^{(j)} = \{\text{annual\_inc}^{(j)}, \text{funded\_amnt}^{(j)}, \text{grade}^{(j)}, \text{installment}^{(j)}, \text{int\_rate}^{(j)}, \cdots\}_{j=1}^{N}$$

and $\boldsymbol{Y}^{(j)} \in \{0, 1\}$, where $\boldsymbol{Y}^{(j)} = 0$ is the dependent variable outcome that shows the loan applicant defaulted on the repayment of the loan and $\boldsymbol{Y}^{(j)} = 1$ is, the opposite outcome.

Some independent variables are catagorical and can be transformed into binary variables. These variables are “term”, “grade”, “verification status” and “home ownership”. The loan “grade” and “home ownership” have more than two levels. There are 7 loan grades from

grade A to grade G, which means that we should create 7 new variables that only consist of two levels, by transforming the "grade" variable into 7 new variables. The transformation is done as follows

$$\text{New variable} = \begin{cases} \text{variable1: if grade is A, then value is } 1, \text{else } 0. \\ \text{variable2: if grade is B, then value is } 1, \text{else } 0. \\ \text{variable3: if grade is C, then value is } 1, \text{else } 0. \\ \cdots \quad \cdots \quad \cdots. \\ \cdots \quad \cdots \quad \cdots. \\ \cdots \quad \cdots \quad \cdots. \\ \text{variable7: if grade is G, then value is } 1, \text{else } 0. \end{cases}$$

All the other catagorical variables were similarly transformed to 2-level binary variables. When the transformation from catagorical variables to two levels binary variables has been successfully done the machine learning models can then be applied. The independent variable called "interest earned", was created to understand the influence of interest to be earned on the loan at marturity stage. However, this variable turns out to have enough information to be used in the prediction of loan default. This variable was calculated as follows

$$\text{interest\_earned} = \text{loan\_amnt} \cdot \text{term} \cdot \text{int\_rate}$$

5.2 Behavioural scorecard

For logistic regression, once the parameters have been estimated, the transfomation to a point based credit scorecard is easily interpretable. Consider the probability of defaulting given as a function of weights of evidence(WOE) to be

$$\mathrm{P}\left(\text{default} = 1|\boldsymbol{X}\right) = \frac{1}{1 + \mathrm{e}^{-\left(\beta_0 + \beta_1 WOE_{X_1} + \beta_2 WOE_{X_2} + \beta_3 WOE_{X_3} + \cdots\right)}}. \tag{5.2.1}$$

In terms of the natural log of the odds equation 5.2.1 can be rewritten as

$$\ln\left(\frac{\mathrm{P}\left(\text{default} = \text{yes}|\boldsymbol{X}\right)}{\mathrm{P}\left(\text{default} = \text{no}|\boldsymbol{X}\right)}\right) = \beta_0 + \beta_1 WOE_{X_1} + \beta_2 WOE_{X_2} + \beta_3 WOE_{X_3} + \cdots.$$

The relationship between credit score and log odds is given by:

$$\text{Credit score} = \text{offset} + \text{factors} \cdot \ln(\text{odds}). \tag{5.2.2}$$

If the rate of change of odds r is specified (e.g., double the odds every 100 points, then r is 2.), equation 5.2.2 becomes

$$\text{Credit score} = \text{offset} + \text{factors} \cdot \ln(r \cdot \text{odds}) - pro, \tag{5.2.3}$$

where pro stands for points at rated odds. Given two credit scores and the corresponding odds, from equation 5.2.2 the offset and factors are easily computed. The point based credit score in terms of weights of evidence of the attributes WOE_{x_i}, logistic co-efficients β_i and intercept β_0 becomes

$$\begin{aligned}\text{Credit score} &= \left(\sum_{i=1}^{N}(WOE_{x_i} \cdot \beta_i) + \beta_0\right) \cdot \text{factors} + \text{offset} \\ &= \left(\sum_{i=1}^{N}\left(WOE_{x_i} \cdot \beta_i + \frac{\beta_0}{N}\right)\right) \cdot \text{factors} + \text{offset}\end{aligned}$$

5.3 Logistic Regression credit scoring model

This study uses the logistic regression statistical technique for credit scoring. This technique is a linear classifier, however different from a linear regression because of its dichotomous outcome variable. The logistic regression estimates the default probability as follows:

$$\mathrm{P}(y = 1|\boldsymbol{X}) = \frac{1}{1 + \mathrm{e}^{-Z}}, \tag{5.3.1}$$

where

$$\begin{aligned}Z &= \theta_0 + \theta_1 \cdot \text{annual\_inc}^{(j)} + \theta_2 \cdot \text{funded\_amnt}^{(j)} + \theta_3 \cdot \text{grade}^{(j)} + \theta_4 \cdot \text{int\_rate}^{(j)} + \cdots \\ &= \theta_0 + \boldsymbol{\Theta}^T \boldsymbol{X}, \quad \boldsymbol{\Theta} = (\theta_1, \theta_2, \theta_3, \cdots, \theta_N)^T\end{aligned}$$

and

$$\mathrm{P}(y = 0|\boldsymbol{X}) = 1 - \mathrm{P}(y = 1|\boldsymbol{X}) \tag{5.3.2}$$

$$= \frac{1}{1 + \mathrm{e}^{\theta_0 + \boldsymbol{\Theta}^T \boldsymbol{X}}}. \tag{5.3.3}$$

The $\boldsymbol{\Theta}$, θ_0 in $Z = \theta_0 + \boldsymbol{\Theta}^{\mathbf{T}}\mathbf{X}$ are, respectively, the parameter vector and scalar intercept. Both the probabilities $\mathrm{P}(y = 1|\boldsymbol{X})$ and $\mathrm{P}(y = 0|\boldsymbol{X})$ in (5.3.1) and (5.3.2), respectively, are bounded between 0 and 1. The parameters of the logistic regression classifier could be estimated using the maximum likelihood procedure, where the log-likelihood fuction is given below as

$$\begin{aligned} LL &= \sum_{j=1}^{N} y^j \log\left(\mathrm{P}\left(y^j = 1|\boldsymbol{x}^j\right)\right) + \left(1 - y^j\right)\log\left(1 - \mathrm{P}\left(y^j = 1|\boldsymbol{x}^j\right)\right) \\ &= \sum_{j=1}^{N} y^j \log\left(\frac{1}{1+\mathrm{e}^{-Z}}\right) - y^j \log\left(\frac{1}{1+\mathrm{e}^{Z}}\right) + \log\left(\frac{1}{1+\mathrm{e}^{Z}}\right) \\ &= \sum_{j=1}^{N} y^j \log\left(\frac{1+\mathrm{e}^{Z}}{1+\mathrm{e}^{-Z}}\right) + \log\left(\frac{1}{1+\mathrm{e}^{Z}}\right) \\ &= \sum_{j=1}^{N} y^j Z - \log\left(1+\mathrm{e}^{Z}\right). \end{aligned}$$

There is no close form solution to maximize the log-likelihood function LL in terms of θ_i. This study uses the vector of partial derivatives called the gradient ascent. The vector of partial derivatives will take the following form on the ith component:

$$\begin{aligned} \frac{\partial LL}{\partial \theta_i} &= \sum_{j=1}^{N}\left(y^j - \frac{\mathrm{e}^{Z}}{1+\mathrm{e}^{Z}}\right)\frac{\partial Z}{\partial \theta_i} \\ &= \sum_{j=1}^{N}\left(y^j - \mathrm{P}\left(y^j = 1|\boldsymbol{x}^j\right)\right)\boldsymbol{X}_i^j, \end{aligned}$$

where the prediction error is represented by the term inside the parenthesis. The prediction error is given by the difference between the observed y^j and its predicted probability $\mathrm{P}(y^j = 1|\boldsymbol{x}^j)$. The standard gradient ascent is used to optimize the logistic regression co-efficients $\boldsymbol{\Theta}$, with the aid of the formula of the derivative of each θ_i. The co-efficients are iteratively updated in the direction of the gradient according to

$$\theta_i \leftarrow \theta_i + \tau \sum_{j=1}^{N}\left(y^j - \mathrm{P}\left(y^j = 1|\boldsymbol{x}^j\right)\right)\boldsymbol{X}_i^j, \tag{5.3.4}$$

where τ is the learning rate.

5.4 Random Forest(RF) credit scoring model

The random forest model constructs a great number of decision trees using the training set and yields the mode of the class and the average prediction of every tree for a classification problem and regression problem, respectively. On the same training set the random forest averages multiple decision trees trained on various parts. The idea behind random forest model is that if one tree produces good results, then a forest of different trees should be better. In other words each tree will likely overfit on some part of the data, many trees built will overfit in different ways, but averaging their results can reduce the amount of overfitting. Random forest model forms a single, strong learner by taking the majority vote from many small and weak decision trees built in parallel.

Let $\boldsymbol{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \cdots, \boldsymbol{S}_k\}$, be the bootstraped sample from $\boldsymbol{D} = \{\boldsymbol{X}_i^j, \boldsymbol{Y}^j\}$. For every $\boldsymbol{S}_i$ measure the level of uncertainty by the entropy$(\boldsymbol{H})$ and the expected reduction to entropy due to sorting of decision attribute$(\boldsymbol{A})$ by the information gain$(\boldsymbol{G})$. The formula for the entropy is shown below as:

$$\boldsymbol{H}(\boldsymbol{S}) = -\rho_{\boldsymbol{f}} \log_2 (\rho_{\boldsymbol{f}}) - \rho_{\boldsymbol{d}} \log_2 (\rho_{\boldsymbol{d}}), \tag{5.4.1}$$

where $\rho_{\boldsymbol{f}}$ represent the proportion of fully paid loans and $\rho_{\boldsymbol{d}}$ represent the proportion of defaulted loans in $\boldsymbol{S}$. Select the optimal features/attributes by calculating the information gain$(\boldsymbol{G})$ per sample $\boldsymbol{S}_t$ on a randonly selected subset of features from $\boldsymbol{X}^j$ at each node of the decision trees. The information gain is calculated as follows:

$$\boldsymbol{G}(\boldsymbol{S}, \boldsymbol{A}) = \boldsymbol{H}(\boldsymbol{S}) - \sum_{a \in values(\boldsymbol{A})} \left(\frac{|\boldsymbol{S}_{\boldsymbol{a}}|}{|\boldsymbol{S}|} \right) \boldsymbol{H}(\boldsymbol{S}_a). \tag{5.4.2}$$

Compute the maximum information gain for each sample $\boldsymbol{S}_t$

$$\psi = \text{argmax}\{\boldsymbol{G}(\boldsymbol{S}, \boldsymbol{A})\}.$$

The final classification for all the trees in the forest is found by averaging the learned distribution $\boldsymbol{P}_t(x|\boldsymbol{S})$ as follows:

$$\boldsymbol{P} = \frac{1}{k} \sum_{t=1}^{k} \boldsymbol{P}_t(x|\boldsymbol{S}).$$

The random forest algorithm works as follows: a boostrap sample $\boldsymbol{S}_i$ is selected from a bootstrap data set $\boldsymbol{S}$, for every tree in forest. The modified decision tree learning algorithm is then used to learn a decision tree. The modified decision tree algorithm is as follows: a

subset of features is selected from the whole set of features $\boldsymbol{X}^j$ and at each node a random subset of the features is picked from that subset and given to the tree. The predictive powers of the random forest is dependent mainly on the following factors: how correlated individual trees are, how each tree performs and the total number of trees.

5.5 Gradient Boosted Decision Trees(GBTs) credit scoring model

The gradient boosting method constructs a more powerful model from combined multiple decision trees. This is an ensemble method that could be used for both regression and classification. Gradient boosting builds the tree that tries to correct the mistakes of the previously built tree, where trees are built in a consecutive manner. If good predictions are provided for only a part of the data by each tree, then more and more trees improve the performance when added iteratively. Gradient boosted model are sensitive to parameter setings, such as the learning rate, which gives the rate at which each tree tries to correct the previous tree's mistakes.

This model does all that by approximating the true function $\boldsymbol{F}(x_i^j)$, in order to minimize the objective non negative convex loss function $\boldsymbol{J}(y_i, \boldsymbol{f}(x_i^j))$, where $\boldsymbol{f}(x_i^j)$ is an approximation of $\boldsymbol{F}(x_i^j)$. The objective function is given below as

$$\boldsymbol{J} = \frac{1}{N}\sum_{i=1}^{N} \log\left(1 + \exp\left(-y_i \boldsymbol{f}(x_i^j)\right)\right).$$

The probability of x_i being correctly classified is given below by

$$\mathrm{P}_i = \frac{1}{\left(1 + \exp\left(-\boldsymbol{f}(x_i^j)\right)\right)}.$$

At each iteration the residual is calculated by $r_i = y_i - \mathrm{P}_i$ and used to build a new tree T_m, then added to the curent ensemble as follows

$$\boldsymbol{f}_m \leftarrow \boldsymbol{f}_{m-1} + \alpha T_m, \quad \text{where } \alpha \text{ is the learning rate.}$$

5.6 Adaptive Boosting(AdaBoost) credit scoring model

Adaptive boosting method assigns weights $\boldsymbol{W}_t$ to each misclassified data point x_i^j based on the level of difficulty faced by previous classifiers. At every iteration a new classifier is trained with the modified training set. At the beginning of this whole process the weights $\boldsymbol{W}_t$ are set to a value 1/N, where N is the number of data points x_i^j. Only the weights of misclassified data points will be updated at each iteration by being multiplied with

$$\delta = (1 - \boldsymbol{s}_\epsilon) / \boldsymbol{s}_\epsilon,$$

where $\boldsymbol{s}_\epsilon$ is the error calculated by adding all the weights of the misclassified points. This error is given below by

$$\boldsymbol{s}_\epsilon = \sum_{i=1}^{N} \boldsymbol{W}_t^{(i)} \boldsymbol{I}^{(i)}, \quad \text{where} \quad \boldsymbol{I} = \begin{cases} 0, & \text{if the data point } x_i \text{ is correctly classified.} \\ 1, & \text{if otherwise.} \end{cases}$$

The updated weights $\boldsymbol{W}_t$ at each iteration becomes

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t \delta.$$

This process stops after a number of certain iteration is reached or if all the data points are correctly classified.

5.7 Stacked Ensemble model

Stacked ensemble method creates a robust model in two-stages. The first stage is the combination of multiple classifiers results and the second stage is using the results from the first stage to train the one last classifier and test the model on the new data set to make the final prediction. This study combines the RF, AdaBoost and GBTs as the first stage classifiers and the second stage classifier as the Logistic regression. The second stage classifier tries to correct the errors made by the first stage classifiers before making predictions using the test data set. The process is demontrated in the diagram below:

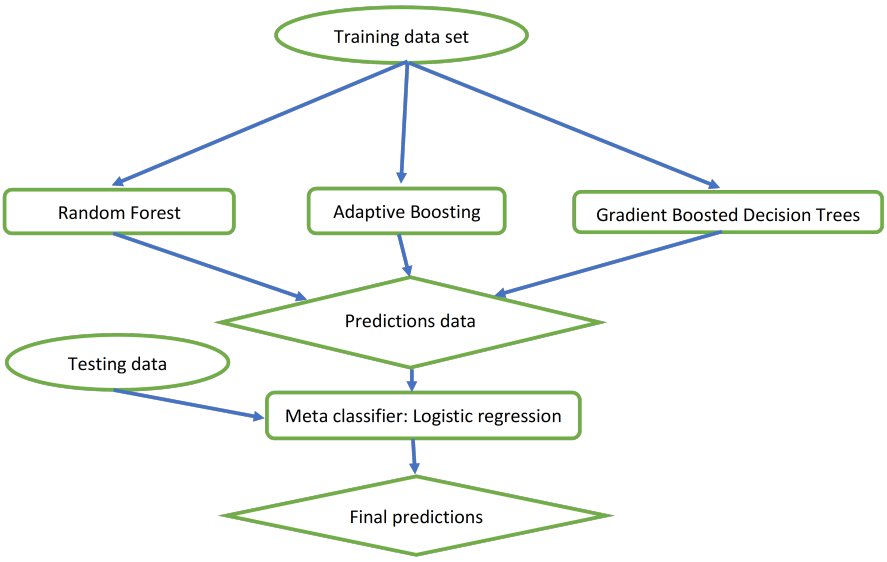


Figure 5.1: Stacked ensemble procedure.

5.8 Reasons for selected methods

The standard credit scoring techniques are built based on the regression techniques (Baesens et al., 2016). Müller et al. (2016) pointed out that random forests is very powerful, robust and always out perform a single decision tree. The gradient boosted decision trees performed better than Adaboost and random forest in the demonstration of social information in P2P lending predictive performance (Niu et al., 2019). Lou and Obukhov (2017) pointed out that in many applications Gradient boosted decision trees have been successful. The gradient boosted decision trees model is more accurate and faster than the random forest (Müller et al., 2016). In the analysis of structure activity relationship of phenol AdaBoost out performed the Artificial Neural Network, Support Vector Machine and K-Nearest Neighbor (Petinrin and Saeed, 2019).

5.9 Predictive Accuracy and Discrimination

This study uses both the confusion matrix or classification table and Receiver Operating Characteristic(ROC) curve to evaluate the performance of a model.

5.9.1 Classification table

Credit scoring models use the classification table as a measure of predictive ability, such as accuaracy and precision. The classification table is constructed based on cross-tabulating the predicted outcomes with the actual observed results. In a case of the two step classifcation problem, a 2-by-2 table of cross classified actual and predicted dichotomous outcomes is constructed. This classification table takes the form below.

		Predicted outcome	
		Fully Paid	**Defaulted**
Actual result	**Fully Paid**	True Positive	False Negative
	Defaulted	False positive	True Negative

Figure 5.2: Classification table

Where *True Positive* mean all the loans that were predicted as "Fully Paid" and were actually fully paid. *False Negative* refers to loans that were predicted as "Defaulted" but were actually fully paid, and *False positive* refers to loans that are predicted to be fully paid, but were actually defaulted. The *True Negative* were the loans predicted as "Defaulted" and were actually default.

The accuracy of the model will be calculated as follows

$$\text{Accuracy} = \frac{\text{True Positive} + \text{True Negative}}{\text{True Positive} + \text{True Negative} + \text{False Negative} + \text{False positive}}.$$

The precision of the model will be calculated as follows

$$\text{Precision} = \frac{\text{True Positive}}{\text{True Positive} + \text{False positive}}. \tag{5.9.1}$$

5.9.2 Discrimination with ROC curves

The ROC curve is also another important method used for performance evaluation of the predictive model. The ROC curve provides a measure of how a predictive model is able to distinguish between the true nagatives and true positives. In a case of loan performance prediction, it is expected that a predictive model should be able to predict a loan "default" as a "default" and a loan that is "fully paid" as "fully paid". The ROC curve method measures the predictive ability of the model by using the area under the curve(AUC). If the AUC is

equal to 0.5 then the model has no predictive ability, but if the AUC is 1.0 then the model has perfect predictive ability. The ROC curve space is divided by the diagonal line to separate the classification results from poor (below the diagonal) to good results(above the diagonal). The diagram below illustrates the ROC curve measure of model prediction ability.

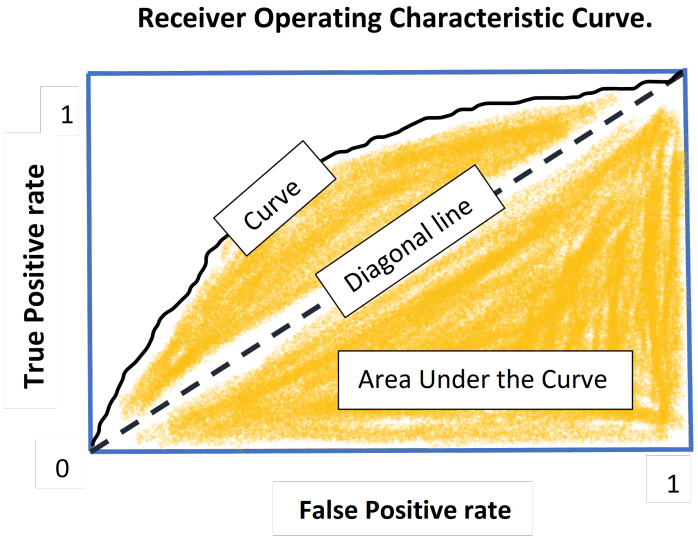


Figure 5.3: The ROC curve model prediction measure

5.10 Model hyperparameter tuning.

This study used the python programming language. The libraries used are: sklearn, numpy, Pandas, seaborn and matplotlib. In sklearn library we used the classifiers namely: LogisticRegression from the linear models package and the RandomForestClassifier, AdaBoostClassifier and GradientBoostingClassifier from the ensemble package. The Pandas library was used for data analysis and data manipulations. The seaborn and matplotlib library were used for data visualization.

The accuracies of machine learning models are mostly dependent on the hyperparameters. These hyperparameters are optimum parameters determined before the application of a machine learning model. The technique that is used to compute the hyperparameter is called grid search. The grid search takes in a set of hyperparameters and allocate optimum values to those hyperparameters. This whole process of tuning parameters saves computing resources, time and effort. The table below contains a set of required hyperparameters for the methods used in this research.

Methods hyper-parameters	**Max features** 1. “sqrt” 2. “log2” 3. “auto”	**Learning rate** $\left[\frac{1}{10}, \frac{1}{100}, \frac{1}{10^3}\right]$	**number of estimators** 10 iterations $[10, 500]$	**Trees max depth** 10 iterations $[1, 10]$
Methods				
Logistic Regression	None	None	None	None
Random Forest	“auto”	None	348	3
Adaptive Boosting	None	None	443	3
Gradient Boosted Trees	None	$\frac{1}{10}$	423	8

Table 5.1: Algorithm hyperparameter grid search.

The logistic regression method has no hyperparameters, due to its mathematical form. However, it has been seen that some parameters play a vital role in the perfomance, namely: *“solver”, “penalty” and “penalty constant C”.* In this study those parameters which optimise the performance of the logistic regression are respectively “liblinear”, “l2”, and 100. There is a need for proper tuning of hyperparameters in the case of random forest, adaptive boosting and gradient boosted trees unlike in the case of logistic regression. This will not be done due to time constraints, since some parameters need to be fixed when varying others.

The hyperparameter namely “*max depth*” which is to reduce the complexity of each tree is kept constant at 3 for both the random forest and adaptive boosting model. For the gradient boosted trees the *“learning rate”* hyperparameter which determines the rate at which each tree learns from the previous trees, was fixed at 0.1. The *“number of estimators”* hyperparameter was calculated by creating a list of 10 numbers and narrowing down to the exact value, by averaging the accuracies found changing the other hyperparameters.

The logistic regression model took 3.5 minutes to find its optimal parameters. The random forest model has three hyperparameters listed on table 5.1, with the *“max depth”* hyperparameter fixed, the model took 431.3 minutes to find the optimal parameters. The adaptive boosting model took 192.0 minutes to only tune the number of estimators needed. The gradient boosted trees model took 1258.1 minutes, which is the longest time taken to find the optimal parameters of all the methods used in this study.

Chapter 6

Results and discussion

This study compares the logistic regression credit scoring method at with the improved machine learning credit scoring methods. The experiments were carried out on the basis of improving the accuracy of the standard credit scoring model. This was achieved by using techniques that utilise information values and multiple correlated variables. The information value that the behavioural credit score carries is higher than that carried by the Fico credit score.

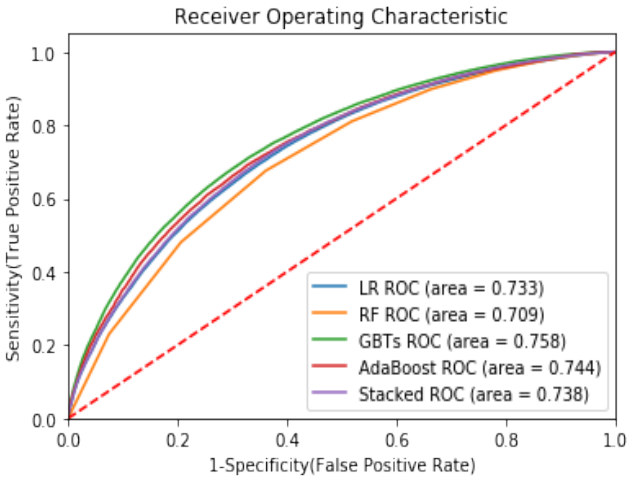
The first experiment compares the impact of two scorecards, developed from different techniques, on the performance of the credit scoring models. The second experiment filters variables that negatively affect the performance of the credit scoring models. The performance of the credit scoring models on both the experiments is measured by the receiver operating characteristic curves, accuracy table and precision.

The training dataset and testing dataset is from the LendingClub cleaned dataset recorded from 2007 to 2018. This LendingClub cleaned dataset is split into the ratio of $3:2$, meaning that the training dataset is 60% and the testing dataset 40% of the whole dataset. To avoid overfitting the 10-fold cross validation of k-fold is used. The out of sample dataset is from the LendingClub dataset recorded from 2019 to the end of the first quarter of 2020.

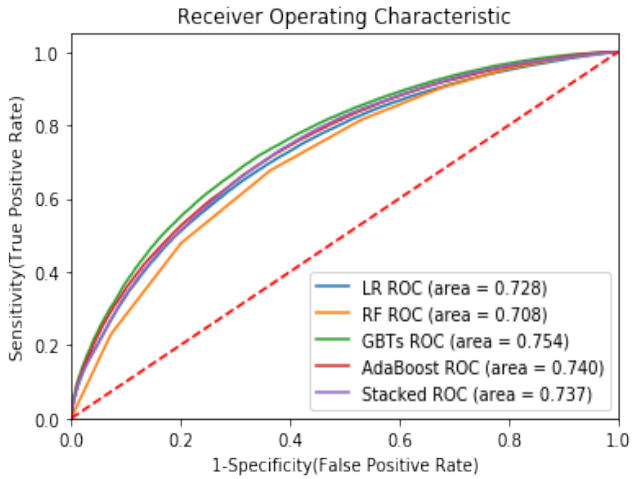
6.1 Comparison between the behavioural credit score and Fico credit score

The behavioural credit score and the Fico credit score are calculated differently. The Fico credit score is from the third party and the formula to calculate the score is unknown to the public. However, the behavioural credit score is calculated using the logistic regression

method. The behavioural credit score information value of 0.880, is higher than the Fico credit score information value of 0.124. This experiment shows the impact of the credit score on the performance of credit scoring models. The diagrams below demonstrate the ROC evaluation of the performance of five credit scoring models used in this study.



(a) The ROC results for behavioural score dataset



(b) The ROC results for Fico score dataset

Figure 6.1: Behavioural score vs Fico score based on cleaned dataset models performance results.

The area values indicated in the figures 6.1a and 6.1b represent the "AUC". Figure 6.1 shows that all the credit scoring models used in this study are able to distinguish between good and bad borrowers. The closer to 1.0 is the area value, the higher the model's prediction capability. The gradient boosted trees(GBTs) model out performed all the models used in this study. However, the behavioural credit score slightly improved the ability of the models to distinguish between good and bad borrowers as seen in figure 6.1a.

Method	Train accuracy	Time taken to train	Test accuracy	Out Of Sample test
LR	0.799	232.00 seconds	0.808	0.802
RF	0.786	344.00 seconds	0.788	0.764
AdaBoost	0.799	258.00 seconds	0.809	0.803
GBTs	0.807	415.00 seconds	0.817	0.810
Stacked Ensemble	0.803	193.00 seconds	0.811	0.806

Table 6.1: Behavioural score based on cleaned dataset model accuracies results.

Method	Train accuracy	Time taken to train	Test accuracy	Out Of Sample test
LR	0.800	232.00 seconds	0.808	0.799
RF	0.791	331.00 seconds	0.791	0.784
AdaBoost	0.805	250.00 seconds	0.813	0.805
GBTs	0.812	399.00 seconds	0.820	0.813
Stacked Ensemble	0.808	190.00 seconds	0.816	0.807

Table 6.2: Fico score based on cleaned dataset model accuracies results.

Although the training dataset is of the same size for both the behavioural score cleaned dataset and Fico score cleaned dataset, tables 6.1 and 6.2 show that the training accuracy for the Fico score cleaned dataset is higher in all the models used in this study. For the Fico score cleaned dataset, the testing accuracies from both the testing cleaned dataset and out of sample cleaned dataset is higher for all the models, except for the logistic regression model.

Precision measures the proportion of the data points for which our model correctly predicted the loan will be fully paid back according to the formula in equation 5.9.1. The aim is to predict correctly that the borrower will not default. To achieve that we have to maximize the precision of the model.

Credit scoring model	Fico score dataset model precision	Behavioural score dataset model precision
Logistic Regression(LR)	0.81**2**1148606	0.81**6**4418608
RandomForest(RF)	0.83**6**0994126	0.83**5**2330566
Adaptive Boosting(AdaBoost)	0.82**2**03598610	0.81**5**1186627
Gradient Boosted Trees(GBTs)	0.8224778071	0.8225362775
Stacked Ensemble	0.82**0**1182830	0.82**1**0989729

Table 6.3: Fico score and behavioural score based on cleaned dataset models precision.

Table 6.3 compares the models precision for both the behavioural score cleaned dataset and Fico score cleaned dataset. The precisions for the prediction of loan default, by the 5 models considered here, in the case of Fico credit score and the behavioural credit score seem to be disagreeing. For every model, the decimal place numbers for which the difference on the precision values occur are indicated in bold font. The LR, GBTs and stacked ensemble models perform better for the behavioural score cleaned dataset than for the Fico score cleaned dataset. Where as the RF and the AdaBoost slightly underperform for the behavioural score cleaned dataset than for Fico score cleaned dataset. The table 6.4 below summarizes the model performances improvement and significance of using behavioural

credit score cleaned dataset over that of Fico credit score cleaned dataset.

Credit scoring model	The difference in model precisions	The percentage of the improvement	The signicant odds predicted correct.
LR	0.004327	1.1847%	$1:100$
RF	-0.000866	None	None
AdaBoost	-0.005241	None	None
GBTs	0.000058	0.6307%	$6:1000$
Stacked Ensemble	0.000981	0.7894%	$8:1000$

Table 6.4: Models performance improvement for behavioural score cleaned dataset.

The difference in model precisions in table 6.4, represent a portion of the cleaned dataset that was incorrectly classified as “fully paid” on the Fico score cleaned dataset. This difference in model precisions is equivalent to a fraction of Fico score cleaned dataset that is “default”, but predicted by the model as “fully paid”. This is very critical information as these differences might have appeared insignificant at a glance but we now can see that it is not. The negative difference in model precisions indicate that there is no improvement, because the credit scoring models underperformed.

The percentage of improvement in table 6.4, is calculated by multiplying the ratio of the difference in model “false positive” predictions from both trials of using behavioural score cleaned dataset and Fico score cleaned dataset to total behavioural score cleaned dataset by 100. The significant odds predicted correct in table 6.4, represent the odds that a model will predict the fully paying loan applicants correct given that the cleaned dataset used has bahavioural score instead of the Fico score. These odds are calculated based on the percentage of model performance improvement. The usage of behavioural credit score maximizes the precision of the LR, GBTs and Stacked Ensemble models.

6.2 Model performance improvement.

In this section we undertake to improve the credit scoring models’ ability to predict required outcome more accurately. Firstly, we conduct an experiment to determine the correlation between the independent variables for the whole dataset. Then, we group the variables with the correlation coefficient(r) of more than $r = 0.8$ and compare the IV of the grouped correlated independent variables. From each group, we consider one independent variable with the highest IV for further calculations. The correlation coefficient $r = 0.8$ is calculated via a series of experiments, observing the effect on the accuracies of all methods used in

this study. For example, when $0.5 \leq r \leq 0.75$ the accuracies of all the methods are lower than when all the independent variables per group are used.

This experiment is conducted for the behavioural credit score dataset only, since this score-card improves the performace of the standard credit scoring model. For $r = 0.8$, only 9 independent variables were removed. These variables are: “avg cur bal”, “bc util”, “funded amnt inv”, “loan amnt”, “num rev tl bal gt 0”, “revol util”, “tot cur bal”, “tot hi cred lim”, “total bc limit”. The perfomance of the methods after these variables were removed are shown below by the ROC, the accuracy table and the precision.

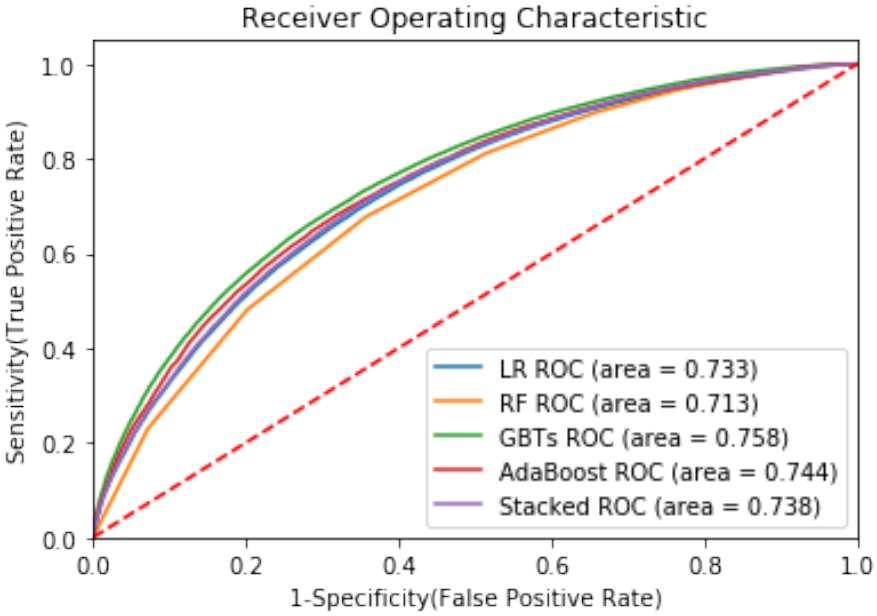


Figure 6.2: The ROC results for uncorrelated variables behavioural score dataset

Figure 6.2 shows that all the models used in this study can still distinguish between good borrowers and bad borrowers. The removal of correlated variables did not affect the ability of model to discriminate. However, the random forest results improved with the area under the curve(AUC) of 0.709 in figure 6.1a increasing slightly to an AUC of 0.713 in figure 6.2. This difference of 0.004 in AUC value is equivalent to 0.4% improvement on the ability of the model to distinguish between good and bad borrowers. The table bellow compares the accuracies of the models before and after the removal of correlated independent variables.

Method	Train accuracy	Time taken to train	Test accuracy	Out Of Sample test
LR	0.799	232.00 seconds	0.808	0.802
RF	0.786	344.00 seconds	0.788	0.764
AdaBoost	0.799	258.00 seconds	0.809	0.803
GBTs	0.807	415.00 seconds	0.817	0.810
Stacked Ensemble	0.803	193.00 seconds	0.811	0.806

Table 6.5: Behavioural score dataset models accuracies results.

Method	Train accuracy	Time taken to train	Test accuracy	Out Of Sample test
LR	0.799	189.00 seconds	0.808	0.803 ↑
RF	0.787 ↑	270.00 seconds	0.789 ↑	0.762 ↓
AdaBoost	0.799	255.00 seconds	0.809	0.803
GBTs	0.808 ↑	351.00 seconds	0.817	0.810
Stacked Ensemble	0.803	150.00 seconds	0.812 ↑	0.806

Table 6.6: Uncorrelated variables behavioural score dataset models accuracies results.

The upward pointing arrow ↑ in table 6.6 represent an improvement achieved by removing correlated variables and the downward pointing arror ↓ indicates the opposite. The random forest model accuracies improved on the training dataset and testing dataset, but slightly underperform on the out sample cleaned dataset testing accuracy. There is a slight improvement on the logistic regression model accuracy, when tested on the out of sample cleaned dataset. However, the other models have maintained the same accuracies when tested on the out of sample cleaned dataset.

To confirm our findings we will also look at the precision of the model before and after the removal of some independent variables. For every model, the numbers for which the difference on the precision values occur will be indicated in bold font, if there is no bold font number then there is no change in model performances.

Credit scoring models	Uncorrelated variables behavioural score dataset model precision	Behavioural score dataset model precision (Before)
Logistic Regression(LR)	0.8164**5**62786	0.8164**4**18608
RandomForest(RF)	0.835**3**525836	0.835**2**330566
Adaptive Boosting(AdaBoost)	0.8151186627	0.8151186627
Gradient Boosted Trees(GBTs)	0.82**3**0314414	0.82**2**5362775
Stacked Ensemble	0.8210**8**29030	0.8210**9**89729

Table 6.7: Behavioural score dataset models precision.

Table 6.7 shows that most of the credit scoring models were slightly affected by the removal of correlated variables, because the change on the precision values occured after the third decimal digit. However, the removal of these variables seem to have improved the precision of all the models except for the adaptive boosted trees model which remained the same and the stacked ensemble model which under performed.

Credit scoring model	The difference in model precisions	The percentage of the improvement	The signicant odds predicted correct.
LR	0.000014	0.0045%	$4:100000$
RF	0.000120	0.0215%	$2:10000$
AdaBoost	0.000000	0.0000%	None
GBTs	0.000495	0.075%	$7:1000$
Stacked Ensemble	-0.000016	None	None

Table 6.8: Models improvement for uncorrelated variables dataset.

The difference in model precisions in table 6.8, represent a portion of the cleaned behavioural score dataset with correlated independent variables that was incorrectly classified as "fully paid". This difference is equivalent to say that for every loan applicants predicted as "fully paid", a fraction of it will be "default" if the independent variables of the cleaned dataset are correlated.

The percentage of improvement in table 6.8, is calculated by multiplying the ratio of the difference in model "false positive" predictions from both trials of cleaned dataset with uncorrelated independent variables and correlated independent variables to total behavioural score cleaned dataset by 100. The significant odds predicted correct in table 6.8, represent the odds that a model will predict the fully paying loan applicants correct given that the cleaned dataset used has uncorrelated independent variables instead of the correlated independent variables. These odds are calculated based on the percentage of model performance improvement. The usage of uncorrelated independent variables maximizes the precision of the LR, RF and GBTs models.

6.3 Conclusion

This work presents the determination of the best way to serve the interests of an investor or lender in lending environment with asymmetric information. In that regard we sought to come up with ways and means of improving current credit worthiness scoring mechanisms in order to minimise chances of defaulting on loan repayment. Four machine learning models

were considered together with the well known LR model. Our, task was to improve on the efficiency of the traditional models. This goal was clearly achieved as evidenced by the results obtained through the machine learning models and validated by the LR method.

Clearly, the use of behavioural score cards in the P2P lending can lead to reduced number of loan defaulters, since it maximises the ability of the standard credit scoring model. This is done through improved model precision as demonstrated in results for the comparison between the behavioural credit score and Fico credit score in section 6.1.

This study also argued that some independent variables negatively affect the performance of the credit scoring model, while others improve it. Since, the removal of 9 independent variables in the experiment done in section 6.2 slightly improved the performance of some models, those variables were negatively affecting the perfomances of those models. Clearly, the removal of those independent variables also changed the ability of the credit scoring model to predict more accurately. This is shown in table 6.8, the odds of the the model to predict correctly improved slightly after the removal of those independent variables that negately affected the model performance.

6.4 Future work

Our study focussed on techniques that are based on the standard credit scoring model. One would like to know if other scorecards could improve the performance as well and how do they compare to the Fico credit score. Most of the datasets in the financial world are not balanced. In the future one might want to look at the impact of the imbalanced classes, and how to reduce model overfitting.

References

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

H. A. Abdou and J. Pointon. Credit scoring and decision making in egyptian public sector banks. *International Journal of Managerial Finance*, 2009.

H. A. Abdou and J. Pointon. Credit scoring, statistical techniques and evaluation criteria: a review of the literature. *Intelligent systems in accounting, finance and management*, 18 (2-3):59–88, 2011.

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

A. Bachmann, A. Becker, D. Buerckner, M. Hilker, F. Kock, M. Lehmann, P. Tiburtius, and B. Funk. Online peer-to-peer lending-a literature review. *Journal of Internet Banking and Commerce*, 16(2):1, 2011.

B. Baesens, D. Roesch, and H. Scheule. *Credit risk analytics: Measurement techniques, applications, and examples in SAS*. John Wiley & Sons, 2016.

F. S. Board. Fintech credit: Market structure, business models and financial stability implications. *Financial Stability Board, Basel*, 2017.

D. Carmichael. Modeling default for peer-to-peer loans. *Available at SSRN 2529240*, 2014.

K. T. Davis and J. Murphy. Peer to peer lending: structures, risks and regulation. *Kevin Davis and Jacob Murphy" Peer to Peer Lending: Structures, Risks and Regulation" JASSA: The Finsia Journal of Applied Finance*, 2016:3–37, 2016.

R. Emekter, Y. Tu, B. Jirasakuldech, and M. Lu. Evaluating credit risk and loan performance in online peer-to-peer (p2p) lending. *Applied Economics*, 47(1):54–70, 2015.

A. Fong. Regulation of peer-to-peer lending in hong kong: state of play. *Law and Financial Markets Review*, 9(4):251–259, 2015.

W. Henley and D. J. Hand. A k-nearest-neighbour classifier for assessing consumer credit risk. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 45(1):77–95, 1996.

L. Larrimore, L. Jiang, J. Larrimore, D. Markowitz, and S. Gorski. Peer to peer lending: The relationship between language features, trustworthiness, and persuasion success. *Journal of Applied Communication Research*, 39(1):19–37, 2011.

Z. Li, X. Yao, Q. Wen, and W. Yang. Prepayment and default of consumer loans in online lending. *Available at SSRN 2740858*, 2016.

Y. Lou and M. Obukhov. Bdt: Gradient boosted decision tables for high accuracy and scoring efficiency. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 1893–1901, 2017.

T. Lynn, J. G. Mooney, P. Rosati, and M. Cummins. Disrupting finance.

A. Mathew. Credit scoring using logistic regression. 2017.

A. Milne and P. Parboteeah. The business models and economics of peer-to-peer lending. 2016.

N. Möllenkamp. Determinants of loan performance in p2p lending. B.S. thesis, University of Twente, 2017.

T. P. Mpofu and M. Mukosera. Credit scoring techniques: a survey. *International Journal of Science and Research (IJSR)*, pages 2319–7064, 2014.

A. C. Müller, S. Guido, et al. *Introduction to machine learning with Python: a guide for data scientists*. " O'Reilly Media, Inc.", 2016.

B. Niu, J. Ren, and X. Li. Credit scoring using machine learning by combing social network information: Evidence from peer-to-peer lending. *Information*, 10(12):397, 2019.

M. Paliwal and U. A. Kumar. Neural networks and statistical techniques: A review of applications. *Expert systems with applications*, 36(1):2–17, 2009.

O. O. Petinrin and F. Saeed. Stacked ensemble for bioactive molecule prediction. *IEEE Access*, 7:153952–153957, 2019.

C. Serrano-Cinca, B. Gutiérrez-Nieto, and L. López-Palacios. Determinants of default in p2p lending. *PloS one*, 10(10), 2015.

N. Siddiqi. *Intelligent credit scoring: Building and implementing better credit risk scorecards*. John Wiley & Sons, 2017.

D. West. Neural network credit scoring models. *Computers & Operations Research*, 27 (11-12):1131–1152, 2000.

B. Xing and T. Marwala. *Smart computing applications in crowdfunding*. CRC Press, 2018.

Appendix A

Variable name	Description
acc open past 24mths	Number of trades opened in past 24 months.
all util	Balance to credit limit on all trades.
annual inc	The self-reported annual income provided by the borrower during registration.
avg cur bal	Average current balance of all accounts.
bc open to buy	Average current balance of all accounts.
bc util	Ratio of total current balance to high credit or credit limit for all bankcard accounts.
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loandivided by the borrower's self-reported monthly income.
earliest cr line	The month the borrower's earliest reported credit line was opened.
emp title	The job title supplied by the Borrower when applying for the loan.
fico range high	The upper boundary range the borrower's FICO at loan origination belongs to.
fico range low	The lower boundary range the borrower's FICO at loan origination belongs to.
funded amnt	The total amount committed to that loan at that point in time.
funded amnt inv	The total amount committed by investors for that loan at that point in time.
grade	LC assigned loan grade.
hardship amount	The interest payment that the borrower has committed to make each month while they are on a hardship plan.
hardship dpd	Account days past due as of the hardship plan start date.
hardship last payment amount	The last payment amount as of the hardship plan start date.
hardship length	The number of months the borrower will make smaller payments than normally obligated due to a hardship plan
hardship loan status	Loan Status as of the hardship plan start date
hardship payoff balance amount	The payoff balance amount as of the hardship plan start date.
hardship reason	Describes the reason the hardship plan was offered.

Variable name	Description
hardship status	Describes if the hardship plan is active, pending, canceled, completed, or broken
hardship type	Describes the hardship plan offering.
home ownership	The home ownership status provided by the borrower during registration or obtained from the credit report. Our values are: RENT, OWN, MORTGAGE, OTHER
inq last 12m	Number of credit inquiries in past 12 months.
installment	The monthly payment owed by the borrower if the loan originates.
int rate	Interest Rate on the loan.
issue d	The month which the loan was funded.
last pymnt d	Last month payment was received.
loan amnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
loan status	Current status of the loan.
mo sin old rev tl op	Months since oldest revolving account opened.
mo sin rcnt rev til op	Months since most recent revolving account opened.
mo sin rcnt til	Months since most recent account opened.
mort acc	Number of mortgage accounts.
mths since recent bc	Months since most recent bankcard account opened.
mths since recent inq	Months since most recent inquiry.
num actv rev tl	Number of currently active revolving trades.
num rev tl bal gt 0	Number of revolving trades with balance > 0.
num tl op past 12m	Number of accounts opened in past 12 months.
open acc 6m	Number of open trades in last 6 months.
open rv 12m	Number of revolving trades opened in past 12 months.
open rv 24m	Number of revolving trades opened in past 24 months.
percent bc gt 75	Percentage of all bankcard accounts $> 75\%$ of limit.
revol util	Revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit.
sub grade	LendingClub assigned loan subgrade.
term	The number of payments on the loan. Values are in months and can be either 36 or 60.
title	The loan title provided by the borrower.
tot cur bal	Total current balance of all accounts.
tot hi cred lim	Total high credit/credit limit.
total bc limit	Total bankcard high credit/credit limit.
total rec int	Interest received to date.
total rec late fee	Late fees received to date.
total rev hi lim	Total revolving high credit/credit limit.
verification status	Indicates if income was verified by LC, not verified, or if the income source was verified.
zip code	The first 3 numbers of the zip code provided by the borrower in the loan application.

Appendix B

Algorithm 1 Logistic Regression algorithm

Require: Training set:$\left\{X^{(n)}, y^{(n)}\right\}_{n=1}^{J}$ and Testing set:$\left\{X^{(n)}, y^{(n)}\right\}_{n=J+1}^{N}$

Ensure: To input Learning rate: τ, Initial parameters: $\Theta^{(0)} = \left[\theta_0^{(0)}, \theta_1^{(0)}, \cdots, \theta_M^{(0)}\right]^T$

Results: Learned parameters: $\Theta = [\theta_0, \theta_1, \cdots, \theta_M]^T$

while True **do**

for $i = 1$ to M **do**

$$\theta_i^{(t)} = \theta_i^{(t-1)} + \tau \sum_{j=1}^{N} \left(y^j - \mathrm{P}\left(y^j = 1 | \boldsymbol{x}^j\right)\right) \boldsymbol{X}_i^j$$

if $\mathrm{is-convergent}$ **then**

break

end if

end for

$t \leftarrow t + 1$

end while

Algorithm 2 Random Forest algorithm

Require: Bootstraped data $\boldsymbol{S} = \{\boldsymbol{S}_1, \boldsymbol{S}_2, \boldsymbol{S}_3, \cdots, \boldsymbol{S}_k\}$ from data set:$\left\{X^{(n)}, y^{(n)}\right\}_{n=1}^{N}$

Ensure: To input the *number of estimators* k, *tree max depth* and *max features*.

Results: Learned distribution $\boldsymbol{P}_t(x|\boldsymbol{S})$

while True **do**

for $\boldsymbol{S}_i$ in $\boldsymbol{S}$ **do**

$$\boldsymbol{G}(\boldsymbol{S}_i, \boldsymbol{A}) = \boldsymbol{H}(\boldsymbol{S}_i) - \sum_{a \in values(\boldsymbol{A})} \left(\frac{|\boldsymbol{S_a}|}{|\boldsymbol{S_i}|}\right) \boldsymbol{H}(\boldsymbol{S}_a)$$

compute:

$$\mathrm{argmax}\{\boldsymbol{G}(\boldsymbol{S}, \boldsymbol{A})\}$$

end for

$$\boldsymbol{P} = \frac{1}{k} \sum_{t=1}^{k} \boldsymbol{P}_t(x|\boldsymbol{S})$$

end while

Algorithm 3 Gradient Boosted Trees algorithm

Require: Training set:$\left\{X^{(n)}, y^{(n)}\right\}_{n=1}^{J}$ and Testing set:$\left\{X^{(n)}, y^{(n)}\right\}_{n=J+1}^{N}$
Ensure: To input Learning rate: α, *number of estimators* k and *tree max depth*.
Results: Learned distribution T_m
while True **do**
for $\boldsymbol{S}$ in Training set **do**

$$\boldsymbol{G}(\boldsymbol{S}, \boldsymbol{A}) = \boldsymbol{H}(\boldsymbol{S}) - \sum_{a \in values(\boldsymbol{A})} \left(\frac{|\boldsymbol{S_a}|}{|\boldsymbol{S}|}\right) \boldsymbol{H}(\boldsymbol{S}_a)$$

compute:
$\operatorname{argmax}\{\boldsymbol{G}(\boldsymbol{S}, \boldsymbol{A})\}$
end for
for $i = 1$ to k **do**

$$\boldsymbol{f}_m \leftarrow \boldsymbol{f}_{m-1} + \alpha T_m$$

if $\boldsymbol{J}$ $\text{is} - \text{convergent}$ **then**
break
end if
end for
$m \leftarrow m + 1$
end while

Algorithm 4 Adaptive Boosting algorithm

Require: Training set:$\left\{X^{(n)}, y^{(n)}\right\}_{n=1}^{J}$ and Testing set:$\left\{X^{(n)}, y^{(n)}\right\}_{n=J+1}^{N}$
Ensure: To input the *number of estimators* k and *tree max depth*.
Results: Learned distribution T_m
while True **do**
for $\boldsymbol{S}$ in Training set **do**

$$\boldsymbol{G}(\boldsymbol{S}, \boldsymbol{A}) = \boldsymbol{H}(\boldsymbol{S}) - \sum_{a \in values(\boldsymbol{A})} \left(\frac{|\boldsymbol{S_a}|}{|\boldsymbol{S}|}\right) \boldsymbol{H}(\boldsymbol{S}_a)$$

compute:

$$\operatorname{argmax}\{\boldsymbol{G}(\boldsymbol{S}, \boldsymbol{A})\}$$

end for
if the data point x_i is correctly classified **then**

$$\boldsymbol{W}_{t+1} = 0.$$

else

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t \left(\frac{1 - \sum_{i=1}^{N} \boldsymbol{W}_t^{(i)}}{\sum_{i=1}^{N} \boldsymbol{W}_t^{(i)}}\right)$$

end if
$t \leftarrow t + 1$
end while