National College of Ireland

Credit-Risk Assessment of Small Business Loans using Naïve Bayes, Decision Tree and Random Forest

MSc Research Project
MSc in FinTech

Anisa Nizar Ahmed
Student ID: x18107656

School of Computing
National College of Ireland

Supervisor: Victor Del Rosal

National College of Ireland

MSc Project Submission Sheet

School of Computing

Student Name:	Anisa Nizar Ahmed
Student ID:	x18107656
Programme:	MSc in FinTech **Year:** 2018-19
Module:	Research Project
Supervisor:	Victor Del Rosal
Submission Due Date:	16/09/2019
Project Title:	Credit-risk Assessment of Small Business loans using Naïve Bayes, Decision Tree and Random Forest
Word Count:	9361 **Page Count** 24

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

Signature:	Anisa Nizar Ahmed
Date:	16/09/2019

PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST

Attach a completed copy of this sheet to each project (including multiple copies)	□
Attach a Moodle submission receipt of the online project submission, to each project (including multiple copies).	□
You must ensure that you retain a HARD COPY of the project, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer.	□

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

Office Use Only	
Signature:	
Date:	
Penalty Applied (if applicable):	

Table of Contents

Credit-Risk Assessment of Small Business Loans using Naïve Bayes, Decision Tree and Random Forest

Anisa Nizar Ahmed
x18107656

Abstract

Small and medium-sized enterprises are the backbone of a country's economy, providing employment to majority of the working population and contributing immensely towards Gross Domestic product (GDP). However, due to the their limited amount of resources and budget, small businesses loans often get rejected by banks who are hesitant to lend due to high credit-risk. The aim of this study is to analyse small business loan applications using machine learning algorithms for identifying factors that lead to high credit-risk. Machine learning can provide a transparent and efficient way of assessing credit-risk than the traditional banking models. Naïve Bayes, Random Forest and Decision Tree are implemented and compared in terms of accuracy. It is seen that interest rate charged by the bank has the highest impact on credit-risk with loans less than 15% annual interest rate having the least credit-risk.

1 Introduction

Small and medium sized enterprises more commonly known as SMEs are the primary source of employment and economic fuel in both developing and developed countries. Due to innovation and disruption in almost every area of technology and business, SMEs are increasing rapidly with their novel and inspiring business models. The European Union (EU) [1] defines SMEs as enterprises with less than 250 employees, less than €50 million annual turnover and less than €43 million annual balance sheet. As depicted in Figure 1, among all the firms in the EU, 99% are SMEs providing employment to about 70% of the total working population (Navaretti, Calzolari and Pozzolo, 2015; Ju and Sohn, 2014). In Ireland, the three major sectors where SMEs engage [2] are technology (28%), life sciences (23%) and fintech (18%), adding up to 70% of the SME industry.

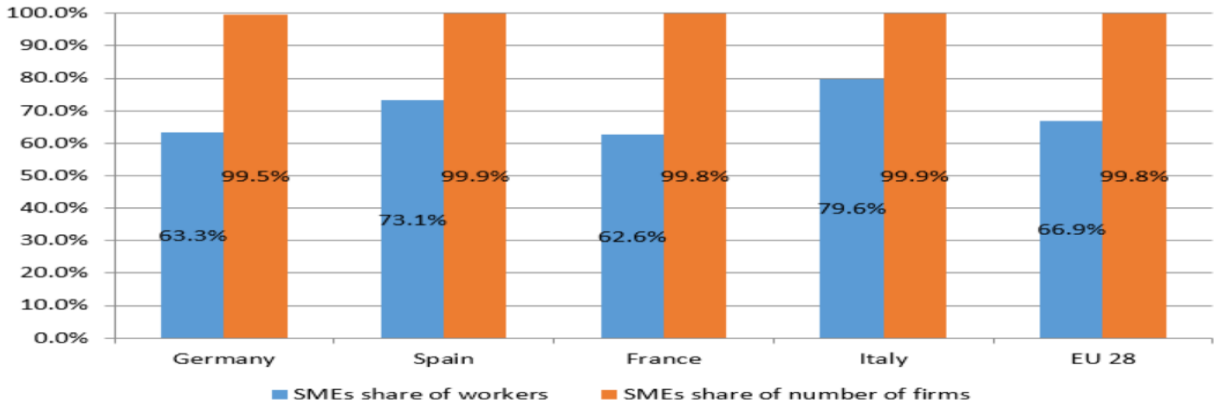

Figure 1: SMEs share of employment and number of firms in Europe.
Source: Navaretti, Calzolari and Pozzolo, 2015

[1] https://ec.europa.eu/growth/smes/business-friendly-environment/sme-definition_en
[2] https://centralbank.ie/docs/default-source/publications/sme-market-reports/sme-market-report-2019.pdf?sfvrsn=9

In spite of steady increase in the number of SMEs formed, there is also a consistent rate of insolvency among them. The EU Commission [3] records that in Ireland, there are three insolvencies among SMEs per day, slightly higher than the average rate in EU. Among the factors that cause insolvencies, access to credit for meeting the short and long term financial needs is the primary reason. A survey carried out by Big Red Cloud [4] of the managers and owners of SMEs in Ireland showed that 58% of respondents had access to credit as their number one challenge while operating a small business. To overcome this challenge, SMEs tap into various external sources of funds like bank loans, government schemes, crowd funding, angel investors, venture capital etc. Among these, bank loans are the preferred source of credit for over 70% of business owners[5].

Nevertheless, acquiring a new loan from a bank requires businesses to have a steady revenue along with outstanding credit history and collateral. Small businesses however are always prone to losses and rarely make profits during the initial years, making them a poor candidate when applying for loans. As there is no stability in their operations or a guarantee of steady income, SME business loans are often prone to high credit-risk i.e. the inability of the borrower to repay the loan amount along with the interest rate charged. In a market economy, a commercial bank requires a high level of effective operations and competitiveness to meet its most important goal of profit maximization (Mileris and Boguslauskas, 2011). This requires the bank to ensure the safety of the raised amount through loans for the timely and complete return of money to its depositors and creditors by accepting high quality and least risky loan portfolios. Thus, banks are hesitant to provide loans to the SMEs as there is minimum or no guarantee of loan repayment leading to high credit-risk. Even in cases when such loans are approved, there are charged with a higher interest rate than a well-established corporate firm. This restricts the SMEs in applying for loans and makes their survival stressful, eventually leading to bankruptcy or insolvency (Bengo and Arena, 2019).

According to the Central Bank of Ireland[6], by the end of 2017, the loan rejection rates were the highest for micro enterprises (less than 10 employees) reaching approximately 30%. This is followed by the total of all SMEs, medium and small with rejection rates of almost 18%, 15% and 10%.

[3] https://dbei.gov.ie/en/Publications/Publication-files/2017-SBA-Fact-Sheet.pdf
[4] https://businessandfinance.com/news/over-half-smes-access-credit-growing-problem/
[5] https://www.irishtimes.com/special-reports/finance-for-smes
[6] https://www.centralbank.ie/docs/default-source/publications/sme-market-reports/sme-market-report-2018.pdf?sfvrsn=6

The data presented by the Central Bank of Ireland in Figure 2 shows that the net lending to all SMEs decreased by 561 million Euros in Q3 2018 whereas, it fell by 920 million compared to previous year.

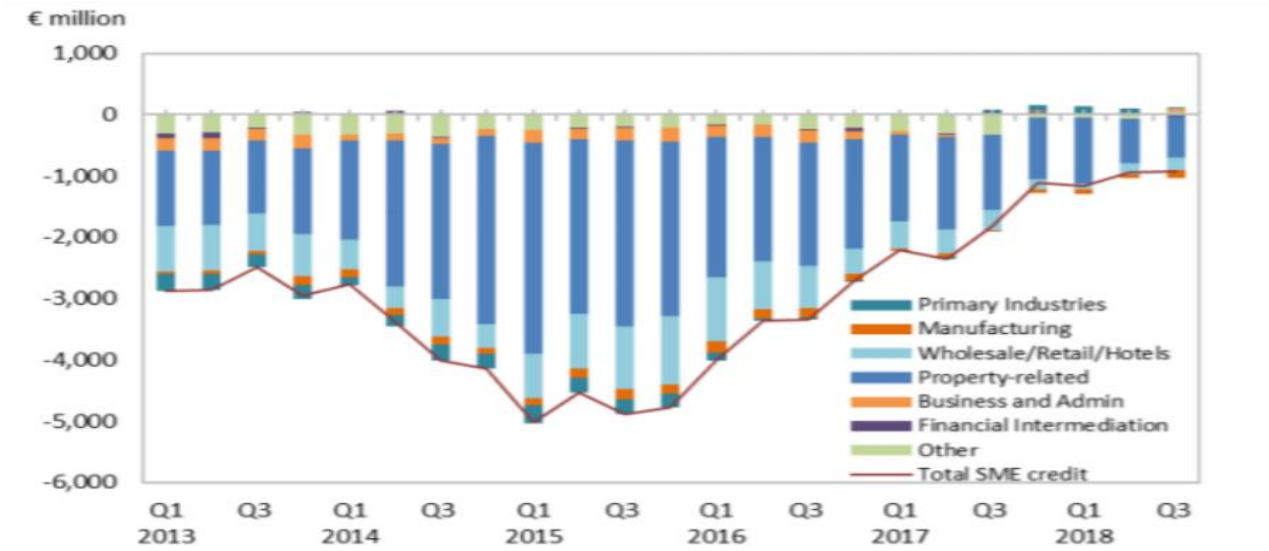

Figure 2: Annual net lending to SMEs in Ireland from 2013-2018
Source: Central Bank of Ireland (2018) [7]

Majority of lending banks have been using traditional methods of evaluating credit-risk in loan applications. Business owners and managers of SMEs are often unaware of such methods and have very little knowledge of what could lead to a high credit-risk while applying for loans. Moreover, banks take a large amount of time in making decisions about a loan application which could even be more than 10 weeks. Thus, there is a need to develop analytical and efficient ways of evaluating credit-risk such that there is minimal use of resources and greater transparency.

This paper aims at developing models using predictive machine learning algorithms namely Naïve Bayes, Random Forest and Decision Trees to evaluate loan data and identify the factors that lead to a good or bad credit loan. It also aims at analyzing these models in terms of accuracy and efficiency such that there is minimal error rate. The Research Questions of the study are:

1. What are the most important factors identified by machine learning algorithms that lead to a high or low credit-risk while evaluating small business loans ?
2. Which model among Naïve Bayes, Random Forest and Decision Tree performs with the highest accuracy?

By achieving the objectives, the paper can contribute towards the application of advanced technology and analytics in the traditional system of finance and lending. It can also provide an effective method for banks to evaluate loans as well as SMEs in analyzing their portfolio and reducing credit-risk.

[7] https://www.centralbank.ie/docs/default-source/statistics/data-and-analysis/credit-and-banking-statistics/business-credit-and-deposits/2018q3_trends_in_sme_and_large_enterprise_credit_and_deposits.pdf?sfvrsn=12

The remainder of the paper is organized as follows. The next section reviews the main theory and previous papers in SME credit-risk assessment. Section 3 describes the data and methodology followed by the design specification implementation of the models in Section 4, and Section 5 evaluates the results and output. Finally, Section 6 gives a detailed discussion on the findings of the experiment along with its limitations which is followed by the conclusion and future work summarizing the entire paper.

2 Related Work

Due to the limited accessibility and transparency of finances in the SME industry, it is often difficult for financial institutions to assess the credit worthiness of such firms (Belas *et al.*, 2018). Previous work in this area have been mostly in proposing different methods and criteria for evaluation of loans. The following section highlight some of the key literature published in the area of SME credit-risk along with their results and limitations.

2.1 Credit-risk

The financial crisis of 2008 had a great impact on the way banks operate. The dynamic economic condition forced the financial institutions to increase regulation and become extremely cautious while evaluating and lending loans (Ferreira *et al.*, 2013). This has required a greater mutual trust between the lender and the borrower i.e., the bank and SME for availing better credit concession. Credit concession is the process of granting monetary value for a guarantee of return of the full amount plus interest rate charged by the lender for a specified term agreed by both the parties (Goncalves *et al.*, 2016). Credit concession always involves some amount of risk on the part of the borrower as the amount is paid in regular installments. When the borrower becomes unable or refuses to repay the money borrowed, it leads to a credit-risk and results in the loss of the lender. Credit-risk may arise due to various reasons like debt, unstable income, lack of proper collateral, poor credit history etc. Additionally, these factors may either be internal or external to the organization. According to Altman and Saunders (1998), the three major external causes that lead to a credit-risk are:

- Increasing rate of bankruptcies
- Higher competition on loan margins
- Asset depreciation

Apart from the external factors, there are a number of internal firm level factors that lead to a high credit-risk. These factors are usually controllable and depend on the type and size of small businesses. Thus, it becomes the primary objective of banks to effectively evaluate these factors and make the best decision on credit concession. Traditionally, banks used various methods of evaluating loan portfolios collectively known as Credit Scoring systems. However, these methods are often complex with little or no transparency leading to unawareness among the SME business owners. This results in a poor relationship between the banks and SMEs due to a lack of trust between them. In their paper, Bengo and Arena (2019) aimed at finding the relationship between SMEs and banks in Italy along with the current practices employed by banks for evaluating credit-risks, and if such practices are beneficial in the long run. Evaluation of data collected from multiple sources show that after the financial crisis, the government of Italy imposed strict and stringent capital requirements by SMEs for credit concession resulting in fewer applications made for loan. It was also evident that most of the loans were applied to larger banks who charged higher interest rates. Following the results, Bengo and Arena (2019) suggest the need for new and improved methods of credit evaluation for assessing risk.

Similarly, a study carried out by Kelly, Brien and Stuart (2014) aimed at analyzing the impact of financial crisis on the SMEs in Ireland and how these organizations could prevent insolvency. Out of the 450,000 firms registered between 1980 and 2012, only 42% were found to be active, indicating a rapid rate of insolvency in Ireland. Furthermore, a survival analysis implemented on the initial data concluded that the unavailability of bank credit and relative stress in macro economy were the primary cause of the high insolvency rate thus hindering the growth of SMEs (Kelly, Brien and Stuart, 2014).

2.2 Credit-Risk assessment of SMEs

Kljucnikov and Belas (2016) examined the critical parameters of credit financing for SMEs in Czech Republic during and after the financial crisis. The study aims at analyzing the behavioral patterns of SMEs such as their relationship with the lending bank and familiarity with finances. Primary data from SMEs were collected via email and telephone after which the responses were categorized into age, gender and qualification level of the firm owner. The size and age of the SME were also taken into account. Statistical differences between the groups were compared and analyzed using an online software through Pearson statistics where the null hypothesis is rejected for a p-value less than 5% (Kljucnikov and Belas, 2016). The study revealed the increasing rate of credit-risk during and after the financial crisis primarily due to the limited amount of banking and financial knowledge posed by the SMEs. They also had very limited knowledge and awareness of the methods employed by banks for assessing credit-risks. Following the study, Kljucnikov and Belas (2016) suggest improved and transparent methods required for credit-risk assessment.

Briozzo, Vigier and Martinez (2016) consider the firm level characteristics like size, age, legal form and industrial sector of the SMEs for evaluating the credit-risk of Argentinian SMEs, a developing economy. The owner's personal characteristics such as age and education is also evaluated to understand if they have an impact on credit decisions. The primary data is obtained through a qualitative approach by surveys and analyzed using an online software (Briozzo, Vigier and Martinez, 2016). The study resulted with the personal characteristics of firm owner influencing the financial decision of the bank and concluded that as the age of the owner increased, there were higher chances of his loan getting approved with a higher grade.

This is further examined by Monika (2016) who writes that the personal characteristics of the firm's owner such as age are also considered while evaluating loan portfolios. A study on 438 SMEs in Slovakia proves that age of the owner has the most significant impact on the interest rate charged by the banks. Similar to Kljucnikov and Belas (2016), Monika (2016) used a qualitative approach for receiving the primary data by emailing a questionnaire to the selected SMEs. Statistical Analysis of the primary data performed using an online software conclude that lower the age of owner, higher is the interest rate charged by the bank as more risk is observed in their loan portfolio**.** Monika (2016) also notes that Slovakia had comparatively higher interest rates than other countries in the EU region due to stringent rules imposed by the government. However, the study by Monika (2016) does not take into account other external factors like social and economic that could have an impact on credit-risk assessment.

In contrast to Monika (2016), Belas *et al.* (2017) argue that the education level of the entrepreneur or firm owner is most important when compared to all other financial as well as non-financial factors. The results also show that financial literacy plays a key role in the relationship between the business and banker, consequently helping in achieving low credit-risk (Belas *et al.*, 2017). Additionally, the family environment within which the owner functions also plays an important role. These results were achieved by performing Structural Analysis Modeling (SEM) using Statistical Package for the Social Sciences (SPSS) and AMOS software on a sample data of 352 SMEs in Czech Republic. Most of the papers mention the limited size of

data as a limitation to their study, paving way for undertaking the research on a larger and diverse data (Kljucnikov and Belas, 2016; Briozzo, Vigier and Martinez, 2016; Monika, 2016).

A Peer-to-Peer platform lending (P2P) channel proposed by Davis, Maddock and Foo (2017) minimizes risks and maximizes profits for SMEs specifically in the FinTech sector of Indonesia. The proposed method uses digital technology for conducting electronic contracting, divisibility and transparency of loan contracts across many lenders and determination of interest rates. Such a platform provides a wider audience and more options for SMEs to borrow credit, avail lower interest rates and flexible payment terms (Davis, Maddock and Foo, 2017). However, the system involves a third-party agent or middlemen who performs all the lending activities and may charge extra fee for the services. Moreover, such lending activities are currently unregulated with no policies in many countries and may be risky to invest in.

Goncalves *et al.* (2016) statistically calculate the weights of multiple factors used for credit-risk assessment. This is performed by developing an idiosyncratic decision support system based on integration of cognitive mapping techniques and Interactive Multiple Criteria Decision Making (TODIM). Goncalves *et al.* (2016) argue that this method performs better in terms of robustness and transparency when compared to other popular algorithms like Linear Discriminant Analysis, Logistics Regression or Multivariate Adaptive Regression Spines. A possible future work mention by Goncalves *et al.* (2016) is the comparison of different credit-risk assessment systems for accuracy and sensitivity. Angilella and Mazzu (2015) classified the SMEs according to internal risks using a multi-criteria approach based on the available information. The internal risks were classified as development, technological, market and production risks. However, these did not have the expected impact on credit decision and further work is required to assess different criteria (Angilella and Mazzu, 2015).

2.3 Machine Learning

A study conducted for analyzing the credit-risk in SME supply chain finance (SCF) in China by Zhu *et al.* (2016a) consist of implementing six types of machine learning algorithms namely individual machine learning (IML) comprising decision tree, ensemble machine learning (EML) using bagging, boosting and random subspace; and integrated ensemble machine learning (IEML) using RS-boosting and multiboosting. The primary data consist of quarterly financial and non-financial information of 48 listed Chinese SMEs in the security market. The algorithms were then compared for accuracy, effectiveness and feasibility resulting in IEML method (RS-boosting) performing with highest accuracy of 85.41%. The least accuracy of 74.80% was observed by EML boosting compared to all other models (Zhu *et al.*, 2016a).

For achieving greater accuracy, Zhu *et al.* (2016b) propose a new machine learning algorithm called Random Subspace-Real AdaBoost (RS-RAB), a type of integrated ensemble machine learning. The same set of data resulted in an accuracy of 86.74%, highest among all the methods used in both papers (Zhu *et al.*, 2016b). Khandani, Kim and Andrew (2010) perform customer credit-risk assessment using a nonlinear nonparametric forecasting model mentioning that machine learning techniques are comparatively more adaptive and suitable to study and analyze the complex and dynamic credit-risk data. The model performs with an accuracy of 85%.

The research gap posed by the above papers can be met by undertaking this study on a larger data using quantitative techniques like machine learning. Most of the previous literature do no take other factors into account like revolving balance and outstanding principal amount which have been considered in this paper. Thus, this study contributes towards improving SME credit-risk assessment methods by proposing the use of automated and accurate machine learning algorithms.

3 Research Methodology

The study of pattern recognition and computational learning theory in artificial intelligence has evolved into Machine Learning which can learn from and make predictions on data (Zhu *et al.*, 2016a). This study implements three classification and prediction models namely Decision Tree, Random Forest and Naïve Bayes to classify the loan application according to different features of the dataset. The primary implementation and analysis is performed using R programming language performed in R Studio. The methodology of the study follows the Cross Industry Process for Data Mining (CRISP-DM) as depicted in Figure 3.

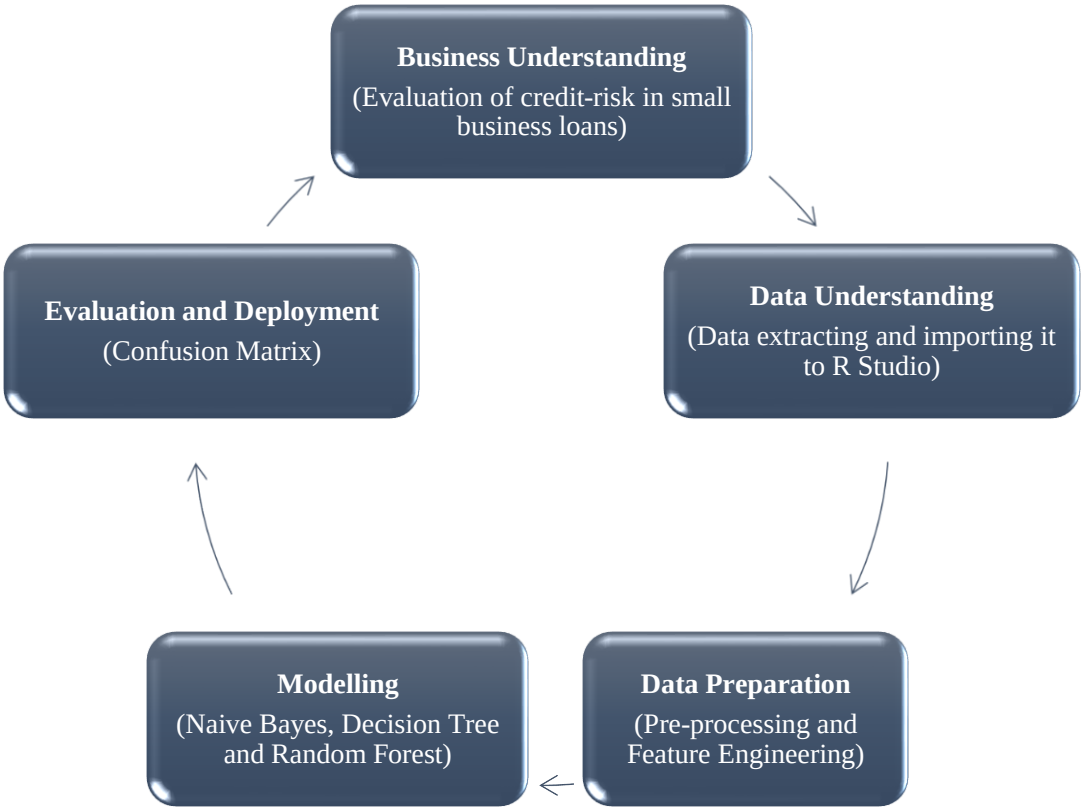

Figure 3: CRISP-DM Framework

SMEs are the most vital industries for a country's economy, their growth and development must be free from hindrance. Currently, availability and access to credit is the major factor limiting the growth of small businesses. Loan applications are being rejecting at an increasing rate by the banks due to high credit-risks and SMEs are unaware of the knowledge required for proper banking relationships. There is a need for transparent, quick and efficient methods of credit-risk assessment by properly classifying the factors and criteria leading to a good or bad loan. To achieve this objective, various attributes of loan application are analyzed and studied followed by the application of machine learning algorithms to identify factors and classify data. Decision Tree, Random Forest and Naïve Bayes are selected for implementation and later compared for accuracy in terms of confusion matrix.

3.1 Data Set

The data set used for analysis is a real world open dataset extracted from Kaggle[8]. The initial dataset consisted of 890,000 observations and 145 variables which were reduced to 99,699 rows and 45 variables to suit the computing power of the device used for analysis. It contains the details of loans granted through 2007-2018 in an Excel sheet. After loading the data onto R Studio, it is analyzed for pre-processing and feature engineering. Due to privacy issues, personal information such as member ID, address line etc. were excluded along with duplicate attributes to create a metadata with 23 attributes and 99,699 rows. Also, "companyage", "homeownership", "verificationstatus", "loanstatus", "disbursementmethod", "Grade" and "Term" are converted into factor variables.

3.2 Data Preparation

The data consist of multiple independent variable both numeric and categorical. For this study, 'Grade' is the dependent or response variable which tells about the loan quality and its associated risk. Banks categorize the loan application into grades based on the level of credit-risk. It is divided into seven levels: A, B, C, D, E, F and G. Loans belonging to level 'A' are the loans with least credit-risk associated with them while loans belonging to level 'G' have the highest credit-risk. The data is then analyzed for missing values and returns a negligible number. The missing values are replaced with '0'.

Variable	Description	Type
loan_amnt	The principal amount borrowed from the bank	Numeric
int_rate	Interest rate charged by the bank per year	Numeric
installment	Amount paid back monthly by the borrower	Numeric
annualrevenue	Annual revenue generated by the business	Numeric
dti	Debt-to-Income ratio	Numeric
deling_2yrs	Number of Dwelling Loan taken	Numeric
Inq_last_6mths	Inquiries done by bank for checking credit reports	Numeric
revol_bal	Revolving balance	Numeric
revol_util	Revolving utilization or Debt-to-limit ratio	Numeric
total_acc	Number of accounts in the past or present	Numeric
out_prncp	Outstanding principal amount	Numeric
total_pymnt	Total amount paid by the borrower	Numeric
total_rec_prncp	Real Estate Contract (REC) principal amount	Numeric
total_rec_late_fee	Late fee charged on REC amount	Numeric
last_pymnt_amnt	Installment amount last paid	Numeric
companyage	Age of the company since registration	Factor with twelve levels
homeownership	Status of accommodating house of the firm's owner	Factor with four levels: Any, Own, Rent, Mortgage,
verificationstatus	Verification status of collateral	Factor with three variables: Source Verified, Verified and Not Verified

[8] https://www.kaggle.com/wendykan/lending-club-loan-data#loan.csv

Grade	Grade of the loan application	Factor with seven levels: A,B,C,D,E,F,G
Term	Repayment term of the loan	Factor with two levels: 36months and 60months
loanstatus	The current status of the loan	Factor with six levels: Charged-off, Current, Fully Paid, In Grace Period, Late (16-30) days, Late (31-120) days.
disbursementmethod	Method of loan repayment	Factor with two levels: Cash and DirectPay

Table 1: Data Features Description

After the pre-processing, the summary of the data is visualized (Table 2) including the statistical calculations of each numeric variable like mean and median.

```
> summary(loandata)
loan_amnt        int_rate      installment      annualrevenue          dti          delinq_2yrs
Min.: 1000    Min.   : 6.00   Min.   :  30.64   Min.   :      0   Min.   :  0.00   Min.   : 0.000
1st Qu.:8000  1st Qu.: 8.81   1st Qu.: 252.98   1st Qu.:  48000   1st Qu.: 11.73   1st Qu.: 0.000
Median :14000 Median :11.80   Median : 380.81   Median :  68717   Median : 17.96   Median : 0.000
Mean   :15936 Mean   :13.00   Mean   : 462.57   Mean   :  83179   Mean   : 19.85   Mean   : 0.231
3rd Qu.:21500 3rd Qu.:16.14   3rd Qu.: 622.68   3rd Qu.: 100000   3rd Qu.: 25.28   3rd Qu.: 0.000
Max.   :40000 Max.   :30.99   Max.   :1618.24   Max.   :9757200   Max.   :999.00   Max.   :24.000

inq_last_6mths   revol_bal         revol_util       total_acc        out_prncp     out_prncp_inv
Min.  :0.0000   Min.   :      0   Min.   :  0.00   Min.   :  2.00   Min.   :    0   Min.   :    0
1st Qu.:0.0000  1st Qu.:   5610   1st Qu.: 24.60   1st Qu.: 14.00   1st Qu.: 7396   1st Qu.: 7391
Median :0.0000  Median :  11179   Median : 42.40   Median : 21.00   Median :12342   Median :12338
Mean   :0.4582  Mean   :  16923   Mean   : 44.28   Mean   : 22.69   Mean   :14722   Mean   :14720
3rd Qu.:1.0000  3rd Qu.:  20539   3rd Qu.: 62.70   3rd Qu.: 29.00   3rd Qu.:19944   3rd Qu.:19944
Max.   :5.0000  Max.   :2358150   Max.   :183.80   Max.   :133.00   Max.   :40000   Max.   :40000

total_pymnt      total_pymnt_inv   total_rec_prncp   total_rec_int    total_rec_late_fee
 Min.   :    0.0   Min.   :    0.0   Min.   :    0.0   Min.   :   0.0   Min.   :  0.00000
 1st Qu.:  644.6   1st Qu.:  644.5   1st Qu.:  399.8   1st Qu.: 170.9   1st Qu.:  0.00000
 Median : 1064.2   Median : 1064.1   Median :  673.1   Median : 331.4   Median :  0.00000
 Mean   : 1648.8   Mean   : 1648.4   Mean   : 1210.4   Mean   : 438.3   Mean   :  0.07248
 3rd Qu.: 1768.7   3rd Qu.: 1768.6   3rd Qu.: 1104.7   3rd Qu.: 597.4   3rd Qu.:  0.00000
 Max.   :41894.4   Max.   :41894.4   Max.   :40000.0   Max.   :3531.3   Max.   :141.60000

 last_pymnt_amnt       companyage     homeownership         verficationstatus
 Min.   :    0.0   10+ years:30059   ANY     :  264   Not Verified   :46306
 1st Qu.:  259.1   < 1 year :12281   MORTGAGE:49302   Source Verified:38003
 Median :  396.0   n/a      : 9193   OWN     :10937   Verified       :15390
 Mean   :  800.3   2 years  : 8400   RENT    :39196
 3rd Qu.:  650.4   3 years  : 7774
 Max.   :41253.5   1 year   : 7333
                   (Other)  :24659
 loanstatus                disbursementmethod Grade           Term
 Charged Off       :   19   Cash     :80141    A:28850   36 months:68423
 Current           :96505   DirectPay:19558    B:28020   60 months:31276
 Fully Paid        : 2426                      C:24454
 In Grace Period   :  314                      D:13289
 Late (16-30 days) :  123                      E: 4859
 Late (31-120 days):  312                      F:  172
                                               G:   55
```

Table 2: Summary of data on R

Summary of data depicted in Table 2 shows that among all the applications, minimum loan amount borrowed was $1000 while the maximum was $40,000. The minimum annual revenue is depicted as '0' which is due to the substitution of missing values as '0' since loans are granted only when there is revenue generated. The age of the companies ranges from 1 year to more than 10 years and the loan repayment term is either 3 years or 5 years. The home ownership tells about the status of the house with the majority of applications belonging to mortgage house. This can also have an effect on credit-risk as there is already a credit remaining to be paid by the borrower. The disbursement method is either cash or direct card payment with the majority of them belonging to cash payment.

4 Design Specification and Implementation

This section gives a brief description of the algorithms implemented in the study.

4.1 Decision Tree

Decision Trees, a supervised machine learning method is one of the most widely and popularly used classification techniques for multiclass variables. It uses branches and leaves to depict the observation of the response or target variable and perform predictive analysis (Pandeya *et al.,* 2017). As depicted in Figure 6, each of the internal node is marked as the input feature and the leaf node is marked with a class or distribution over the classes (Lee *et al.,* 2006). The final value of the response variable is marked by the terminal node whereas the branches between the nodes contain the values that the attributes can have (Pandey *et al.,* 2017). The algorithm predicts the value of the target variable by learning from the training data. For this study, Decision Tree is selected as it is easy to understand and often follows the same method of making a decision as humans. Moreover, decision tree works well with categorical features as present in the dataset.

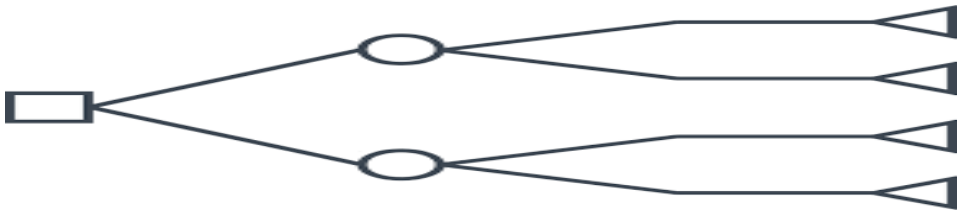

Figure 4: Diagrammatic representation of Decision Tree
Root nodes – represented by square
Branch nodes – represented by circle
Terminal nodes – represented by triangle

4.2 Naïve Bayes

The Bayesian Theorem is the basis for Naïve Bayes algorithm which works well even in a small sample data with multiclass prediction as required in this study. The Bayesian network is an annotated acyclic graph that represents a joint probability distribution over a set of random variables (Pandey *et al.,* 2017). Naïve Bayes also does not incur the problem of overfitting as it adapts to changes in the data. It also works well with text classification and requires less model training time.

4.3 Random Forest

Random Forest is a flexible machine learning algorithm which is used for both classification and regression problems. It produces accurate results even without hyper-parameter tuning (Oughali, Bahloul and El

Rahman, 2019). The model builds and merges multiple decision trees to make an accurate and stable prediction overcoming the problem of overfitting. Moreover, Random Forest also works well with categorical data as required in this study (Oughali, Bahloul and El Rahman, 2019).

5 Results and Evaluation

The dataset is split into 80% training and 20% test data. The models are then fitted into the data, factors identified and predicted the Grade of a loan. After this, all three models are compared in terms of accuracy.

5.1 Naïve Bayes

After splitting the data into 80% training and 20% test, Naïve Bayes is fitted into the training data.

```
> model1 <- naive_bayes(GRADE ~ ., data = train1)
> model1
==================== Naive Bayes ====================
Call:
naive_bayes.formula(formula = GRADE ~ ., data = train1)

A priori probabilities:

           1            2            3            4            5            6            7
0.2892731830 0.2814786967 0.2444736842 0.1338220551 0.0486716792 0.0017418546 0.0005388471

Tables:

loan_amnt         1         2         3         4         5         6         7
     mean 15785.081 16493.354 15984.065 15722.900 14019.130 15707.194 11462.791
     sd   10474.515 10390.725  9924.864  9381.472  9442.405 11182.706 11757.956


int_rate           1           2           3           4           5           6           7
    mean  7.55424363 11.38703633 15.12348506 19.63365671 25.11685376 29.54100719 30.84813953
    sd    0.87977457  0.92325358  1.18619407  1.38719806  1.56394507  0.80471824  0.06630578


installment        1        2        3        4        5        6        7
       mean 460.1252 456.0441 461.1332 479.2536 481.5551 530.5177 410.4628
       sd   302.0637 276.3568 273.6076 278.5872 317.8693 372.7194 411.5255


annualrevenue         1         2         3         4         5         6         7
         mean  90068.62  83166.96  79160.42  77396.18  74266.86  68141.47  53798.05
         sd   107620.37  86946.19 117403.14 113203.88  57557.10  61958.06  31587.09


dti           1        2        3        4        5        6        7
  mean 17.50281 19.51231 20.87287 22.27030 23.49732 24.59460 20.20209
  sd   16.66849 17.86726 17.52271 20.33370 28.09128 19.31235 12.96426

# ... and 17 more tables
```

Figure 5: Naïve Bayes Model on Training Data

As seen in Figure 5, the model shows that the highest number of loans i.e., 28.92% are categorized into ‘A’ grade which are the best loan applications. This is closely followed by ‘B’ with 28.14% loan applications. The least number of loans belong to grade ‘G’ with less than 0.05% of the total loan applications. The model also depicts the mean value and standard deviation of the features according to their grades. The amount of loan borrowed, and installment amount does not show a pattern among the grade categories.

However, there seems to be a strong correlation between the grade and interest rates; as the grade of the loan decreases, the interest rate increases gradually. There is a drastic difference of interest rate charged between grade 'A' and grade 'G'. A similar relationship is seen between the annual revenue of the borrower and the grade of the loan. Applicants belonging to grade 'A' have an average revenue of $90,068 per annum while those belonging to grade 'G' have $53,798 per annum. Debt-to-Income (DTI) ratio also plays a key role and shows a negative correlation with lower grade applications having higher DTI up to grade 'F' after which there is a drop in DTI. The correlation coefficient of grade and interest rate is calculated returning 0.973, depicting a strong positive correlation between the two variables.

Grade	Loan Amount	Annual Interest Rate	Annual Revenue	Debt-to-Income Ratio	Installment paid per month
A	15785.081	7.55%	90068.62	17.503	460.1252
B	16493.354	11.39%	83166.96	19.512	456.0441
C	15984.065	15.12%	79160.42	20.873	461.1332

Table 3: Mean Values of Selected Variables in the top three Grade

As depicted in Table 3, the top three grades contain similar loan amounts and installment amount. However, firms with annual revenue of greater than $85,000 falls into grade A and are charged with the lowest interest rate of 7.55% average. There is a considerable amount of difference in interest rate charged as the annual revenue decreases and DTI ratio increases. For this bank, it is safe to say that firms with annual revenue of more than $80,000 and DTI ratio of less than 20% are likely to fall into grade A with minimum credit-risk.

The model is then applied to the test data for predicting loan grade according to the other independent variables.

```
> q <- predict(model1,test1)
> q
   [1] 3 4 3 3 4 5 1 5 4 3 3 3 2 3 2 4 3 2 5 3 2 4 3 3 4 3 3 2 3 3 4 7 2 3 5 3 3 3 5 3 4 3 3 2 5 4 2 2 1
  [50] 2 2 3 4 3 3 1 2 1 1 1 2 3 3 2 3 4 3 2 1 2 3 1 2 4 3 4 7 2 2 2 3 2 3 2 5 2 2 1 1 2 1 2 3 4 1 1 2 4
  [99] 2 2 1 3 3 2 1 1 1 2 3 2 5 1 3 2 2 1 1 3 4 3 3 3 2 2 1 3 2 1 3 1 2 1 1 3 2 3 3 2 1 4 1 2 2 3 3 1 2
 [148] 2 5 2 1 2 2 1 1 2 1 2 2 1 2 1 2 1 1 1 3 1 1 2 1 2 2 5 1 1 2 5 3 3 2 2 3 4 4 3 3 1 2 1 2 1 1 1 1 2
 [197] 1 3 4 1 3 1 2 4 3 1 1 4 3 3 1 2 3 2 4 2 2 1 1 1 2 1 7 2 3 1 2 3 2 3 5 3 3 2 3 4 1 1 2 4 1 4 7 2 1
 [246] 1 4 1 4 1 1 1 1 1 2 1 1 2 2 3 1 5 3 5 2 1 3 1 1 2 1 2 2 5 2 1 2 2 3 2 1 3 5 3 3 1 3 1 1 1 2 3 2 2
 [295] 1 2 1 3 2 2 2 7 1 3 2 2 3 3 1 1 4 5 1 3 2 3 1 3 7 3 7 2 1 2 1 1 1 1 1 4 3 2 4 1 1 1 1 1 2 3 1 1 4
 [344] 1 3 1 3 1 3 2 3 3 1 1 5 4 1 3 5 1 2 7 3 2 3 3 1 1 7 3 2 7 2 2 1 2 2 1 3 3 2 2 1 1 2 2 3 2 1 3 1 2
 [393] 2 1 3 3 1 1 7 3 1 1 2 7 3 2 3 2 1 1 3 1 3 3 3 1 4 4 1 1 3 4 3 1 2 1 4 2 2 1 1 3 4 1 1 1 3 7 1 1 1
 [442] 2 1 2 3 1 2 5 1 1 1 1 4 2 3 3 1 2 2 1 3 2 3 5 1 4 3 1 1 1 4 3 1 2 3 1 1 2 2 3 4 1 1 1 1 2 2 4 1 3
 [491] 3 3 2 1 2 1 3 1 2 2 2 2 3 1 2 7 2 2 1 7 2 1 7 2 1 2 2 1 3 1 2 1 3 3 3 4 3 3 1 3 3 2 3 1 2 3 3 1 3
 [540] 1 4 1 1 2 1 3 2 1 2 1 3 3 2 7 1 2 3 3 2 1 1 1 1 3 4 2 3 1 3 3 2 3 1 1 3 1 4 1 3 1 1 2 3 2 2 1 5 5
 [589] 1 4 2 3 2 1 1 5 2 2 1 5 1 2 2 1 2 1 3 2 3 1 2 2 3 3 2 3 2 3 1 1 1 3 4 2 1 2 3 1 4 4 1 5 3 5 3 5 3
 [638] 2 3 1 2 1 1 3 4 1 1 2 3 4 1 3 3 1 4 2 1 4 5 3 2 1 2 3 3 4 1 4 1 3 1 2 1 3 1 4 1 4 3 3 7 3 1 7 3 1
 [687] 3 2 4 2 2 3 3 1 2 4 2 2 1 3 2 2 4 2 1 2 3 4 1 3 1 1 1 3 2 2 5 2 2 3 1 2 1 2 4 1 2 1 1 1 2 7 3 1 1
 [736] 1 2 4 4 3 4 3 2 1 2 2 3 1 2 1 1 2 2 1 5 3 7 2 1 2 1 1 2 2 1 2 1 1 2 3 1 2 3 4 7 2 3 5 2 4 1 2 3 3
 [785] 2 1 1 3 1 1 1 2 2 3 2 3 1 4 2 7 1 1 5 4 1 2 7 4 7 4 5 2 4 2 2 5 7 4 4 1 4 2 1 4 3 3 1 1 2 1 2 2 2
 [834] 2 2 3 3 1 1 5 2 2 2 1 2 1 2 1 3 1 5 3 4 1 2 1 1 2 3 2 3 2 3 1 2 2 2 1 2 1 2 3 4 3 1 2 2 2 3 2 2 1
 [883] 4 2 1 3 4 2 1 2 4 1 1 2 2 2 3 2 5 1 1 1 2 3 1 3 7 3 4 1 2 2 2 3 2 2 4 2 4 4 1 3 1 7 4 3 4 1 1 2 2
 [932] 2 2 7 2 1 4 1 2 1 1 1 2 1 3 2 1 5 3 3 1 3 1 2 2 1 1 3 3 1 3 3 2 7 3 3 3 1 4 4 3 2 3 4 1 1 3 2 3 3
 [981] 5 1 3 2 2 2 2 1 2 1 4 3 2 3 4 2 2 2 1 3
 [ reached getOption("max.print") -- omitted 18899 entries ]
Levels: 1 2 3 4 5 6 7
```

Figure 6 (a): Naïve Bayes Model on Test Data for Prediction

```
> confusionMatrix(table(q, test1$GRADE))
Confusion Matrix and Statistics
```

q	1	2	3	4	5	6	7
1	5618	4	1	0	0	0	0
2	58	5329	79	0	1	0	0
3	7	87	4686	23	1	0	0
4	4	4	33	2428	16	0	0
5	1	3	8	35	836	22	1
6	0	0	0	1	3	8	0
7	78	131	138	123	118	3	11

Table 4(a): Confusion Matrix

```
Overall Statistics
               Accuracy : 0.9506
                 95% CI : (0.9475, 0.9536)
    No Information Rate : 0.2898
    P-Value [Acc > NIR] : < 2.2e-16
                  Kappa : 0.9352
Mcnemar's Test P-Value : NA
Statistics by Class:
```

	Class:1	Class:2	Class:3	Class:4	Class:5	Class:6	Class:7
Sensitivity	0.9743	0.9588	0.9476	0.9303	0.85744	0.242424	0.9166667
Specificity	0.9996	0.9904	0.9921	0.9967	0.99630	0.999799	0.9702821
Pos Pred Value	0.9991	0.9748	0.9754	0.9771	0.92274	0.666667	0.0182724
Neg Pred Value	0.9896	0.9841	0.9828	0.9895	0.99268	0.998743	0.9999482
Prevalence	0.2898	0.2793	0.2485	0.1312	0.04900	0.001658	0.0006030
Detection Rate	0.2823	0.2678	0.2355	0.1220	0.04201	0.000402	0.0005528
Detection Prevalence	0.2826	0.2747	0.2414	0.1249	0.04553	0.000603	0.0302528
Balanced Accuracy	0.9870	0.9746	0.9699	0.9635	0.92687	0.621111	0.9434744

Table 4(b): Confusion Matrix

The accuracy of the model is calculated using the Confusion Matrix that returns an accuracy of 95.06% which is higher than the accuracy achieved in previous works. Naïve Bayes performs with a very good accuracy in predicting the loan quality, given the information.

5.2 Random Forest

The data is again split into 70% training data and 20% test data for the implementation of Random Forest model. It is fitted to the training data with 'Grade' as the response variable. The model performs with an error rate of 0.17% as shown in Table 5(a). The most important factors are also calculated according to their importance value (Table 5 (b) and 5 (c)) by using Random Forest algorithm which shows that interest rate is the most important variable while deciding the grade of the loan.

```
> rf ← randomForest(Grade~., data = training4)
> print(rf)
Call:
  randomForest(formula = Grade ~ ., data = train
                Type of random forest: classification
                      Number of trees: 500
   No. of variables tried at each split: 4
           OOB estimate of error rate: 0.17%
```

Confusion matrix:

	A	B	C	D	E	F	G	class.error
A	20163	0	0	0	0	0	0	0.000000e+00
B	1	19707	0	0	0	0	0	5.074082e-05
C	3	0	16997	0	0	0	0	1.764706e-04
D	0	0	0	9324	3	0	0	3.216468e-04
E	1	0	0	11	3359	0	0	3.559775e-03
F	0	0	0	0	72	51	0	5.853659e-01
G	0	0	0	0	21	9	8	7.894737e-01

Table 5(a): Confusion Matrix – Train Data

```
>  importance (rf)
                                        MeanDecreaseGini
loan_amnt                                     1046 . 569141
int_rate                                     33029 . 401458
installment                                   1539 . 376981
annualrevenue                                  228 . 805143
dti                                            258 . 001481
deling_2yrs                                     62 . 757075
inq_last_6mths                                  79 . 332607
revol_bal                                      254 . 090708
revol_util                                     577 . 81
total_acc                                      173 . 167436
out_prncp                                     1342 . 121394
out_prncp_inv                                 1279 . 532835
total_pymnt                                   1246 . 834430
total_pymnt_inv                               1302 . 121839
total_rec_prncp                               3261 . 015223
total_rec_int                                 3995 . 021781
total_rec_late_fee                               4 . 176897
last_pymnt_amnt                               1295 . 819925
companyage                                     282 . 358223
homeownership                                   57 . 301909
verificationstatus                              89 . 678696
loanstatus                                      40 . 376250
Disbursementmethod                             763 . 402692
Term                                           559 . 530297
```

Table 5(b): Most Important Variables

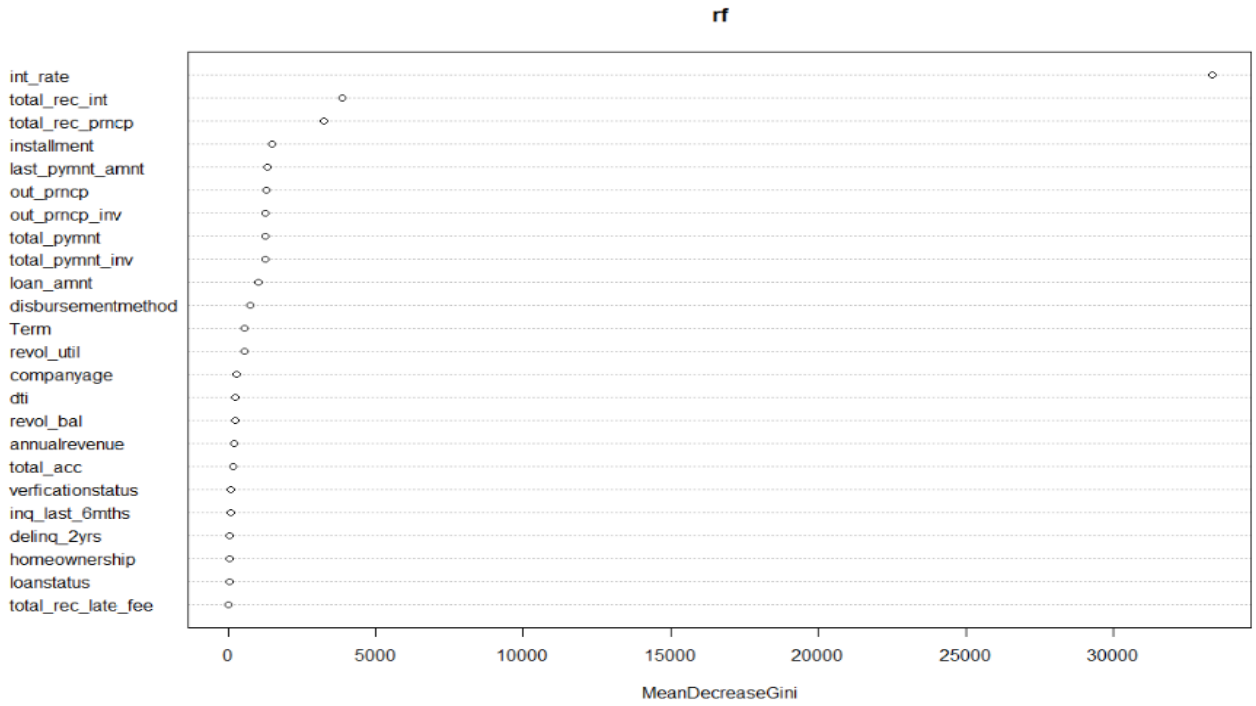

Table 5(c): Variable Importance Graph

The model is implemented to the test data for predicting the loan grade as shown in Figure 7.

```
> p2 <- predict(rf, newdata = test4, type = "class")
> p2
```

1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
C	D	C	C	D	E	D	A	E	B	B	A	D	C	A	C	B	C	B	C
21	22	23	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40
B	C	B	D	C	B	C	C	E	C	B	D	C	C	D	C	C	C	B	C
41	42	43	44	45	46	47	48	49	50	51	52	53	54	55	56	57	58	59	60
C	D	A	B	B	D	B	C	C	E	C	C	C	B	E	D	C	C	B	E
61	62	63	64	65	66	67	68	69	70	71	72	73	74	75	76	77	78	79	80
C	C	C	B	B	B	E	D	B	B	B	B	D	D	B	D	A	B	A	A
81	82	83	84	85	86	87	88	89	90	91	92	93	94	95	96	97	98	99	100
B	C	C	D	C	C	A	A	B	A	A	E	A	B	C	C	B	C	A	C
101	102	103	104	105	106	107	108	109	110	111	112	113	114	115	116	117	118	119	120
D	C	A	B	B	A	B	D	C	A	C	B	D	C	B	C	E	C	C	A
121	122	123	124	125	126	127	128	129	130	131	132	133	134	135	136	137	138	139	140
D	A	B	A	B	B	C	C	B	D	C	B	B	E	B	B	B	A	B	A
141	142	143	144	145	146	147	148	149	150	151	152	153	154	155	156	157	158	159	160
B	A	B	A	D	A	A	B	C	D	A	A	A	C	B	E	B	B	B	A
161	162	163	164	165	166	167	168	169	170	171	172	173	174	175	176	177	178	179	180
C	C	A	C	B	B	A	A	A	D	B	C	B	E	A	C	A	B	B	A
181	182	183	184	185	186	187	188	189	190	191	192	193	194	195	196	197	198	199	200
A	C	D	C	D	C	B	A	B	A	C	B	A	C	A	B	A	A	C	B

Figure 7: Predicted Grades by Random Forest

```
> confusionMatrix(table(p2, test4$Grade))
Confusion Matrix and Staistics
```

p2	A	B	C	D	E	F	G
A	8687	0	1	0	1	0	0
B	0	8312	0	0	0	0	0
C	0	0	7453	1	0	0	0
D	0	0	0	3959	4	1	0
E	0	0	0	2	1483	27	11
F	0	0	0	0	0	21	4
G	0	0	0	0	0	0	2

Table 6(a): Confusion Matrix – Random Forest

```
Overall Statistics
               Accuracy: 0.9983
                 95% CI: (0.9977, 0.9987)
    No Information Rate: 0.2899
    P-Value [Acc > NIR]: < 2.2e-16
                  Kappa: 0.9977
Mcnemar's Test P-Value: NA
Statistics by Class:
```

	Class:A	Class:B	Class:C	Class:D	Class:E	Class:F	Class:G
Sensitivity	1.0000	1.0000	0.9999	0.9992	0.99664	0.4285714	1.176e-01
Specificity	0.9999	1.0000	1.0000	0.9998	0.99860	0.9998663	1.000e+00
Pos Pred Value	0.9998	1.0000	0.9999	0.9987	0.97374	0.8400000	1.000e+00
Neg Pred Value	1.0000	1.0000	1.0000	0.9999	0.99982	0.9990649	9.995e-01
Prevalence	0.2899	0.2774	0.2487	0.1322	0.04965	0.0016350	5.673e-04
Detection Rate	0.2899	0.2774	0.2487	0.1321	0.04948	0.0007007	6.674e-05
Detection Prevalence	0.2899	0.2774	0.2487	0.1323	0.05082	0.0008342	6.674e-05
Balanced Accuracy	1.0000	1.0000	0.9999	0.9995	0.99762	0.7142189	5.588e-01

Table 6(b): Confusion Matrix – Random Forest

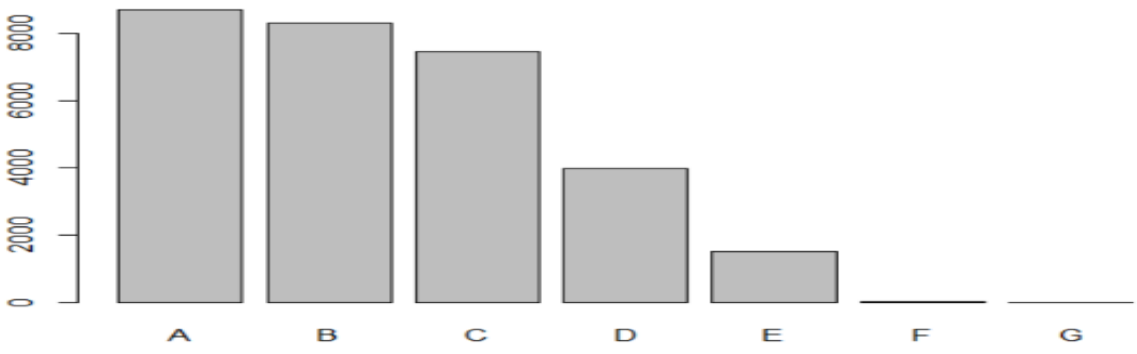

Figure 8: Prediction Graph

The confusion matrix (Table 6(a) and 6(b)) calculates the accuracy of Random Forest model to be 99.83% which is a nearly perfect prediction by the algorithm. The prediction graph shows a similar pattern as seen in the summary of the data with the majority of loans belonging to grade A. Random Forest performs better than Naïve Bayes in predicting the loans.

5.3 Decision Tree

Decision Tree is implemented to 70% training data, first with Grade as the response variable and all other attributes as independent variables.

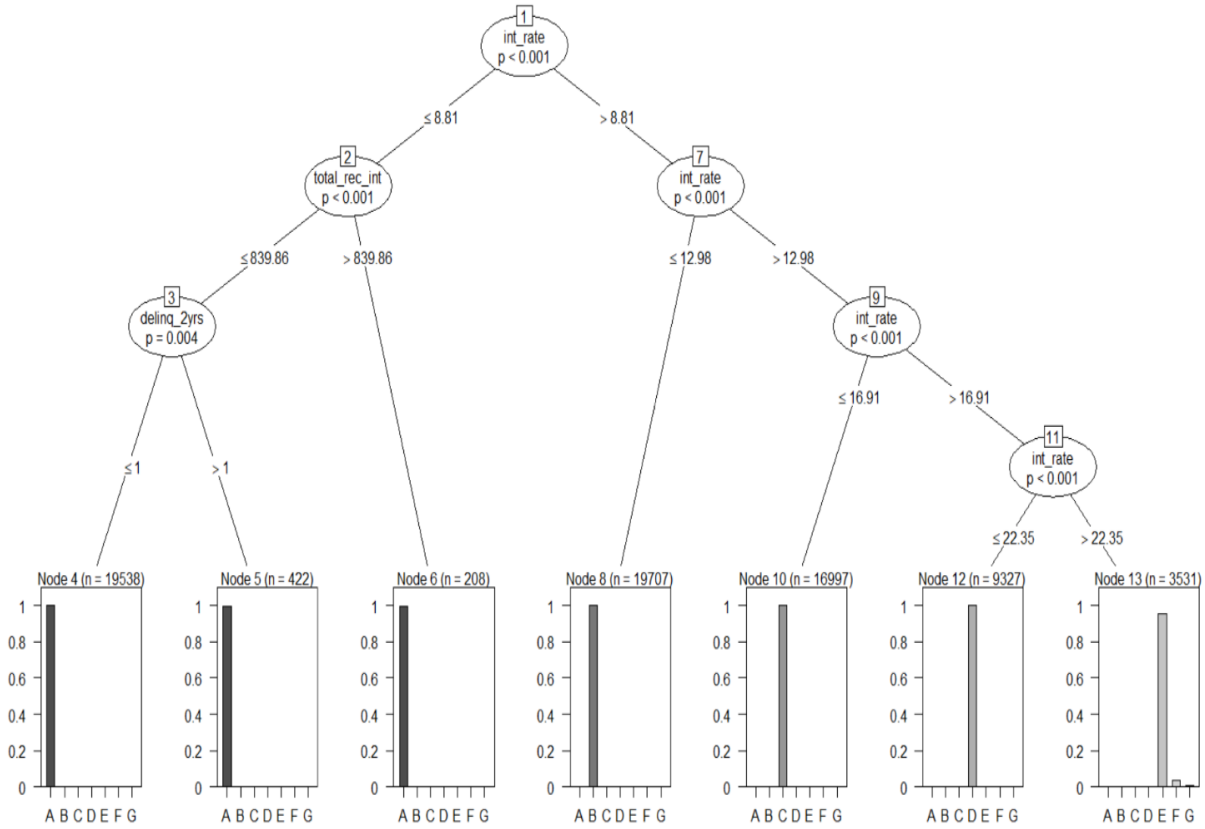

Figure 9: Decision Tree with all Independent Variables

Decision Tree identifies interest rate as the most important factor placing it on the root node. Similar to Naïve Bayes, loans with annual interest rate of less than 8.81% are categorized as grade 'A' loans while those with annual interest rate less than 12.98% belong to grade 'B'. Figure 9 further shows that 'total_rec_int' and Dwelling Loan does not have a great impact on the loan grades. Interest rates as high as 22.35% per annum are charged on the bad loans belonging to grade 'E'. The model is then fitted to the test data, which performs with an accuracy of 99.99% as shown in Table 7.

testpred	A	B	C	D	E	F	G
A	8687	0	1	0	1	0	0
B	0	8312	0	0	0	0	0
C	0	0	7453	0	0	0	0
D	0	0	0	3962	0	0	0
E	0	0	0	0	1487	49	17
F	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0

```
> 1-sum(diag(tab))/sum(tab)
[1] 0.002269011
```

Table 7: Confusion Matrix – Decision Tree with all variables.

Due to the high accuracy achieved, the model is trained again with only selected independent variables: loan amount, annual revenue, DTI, revolving balance, age of the company, verification status, disbursement method and term of repayment.

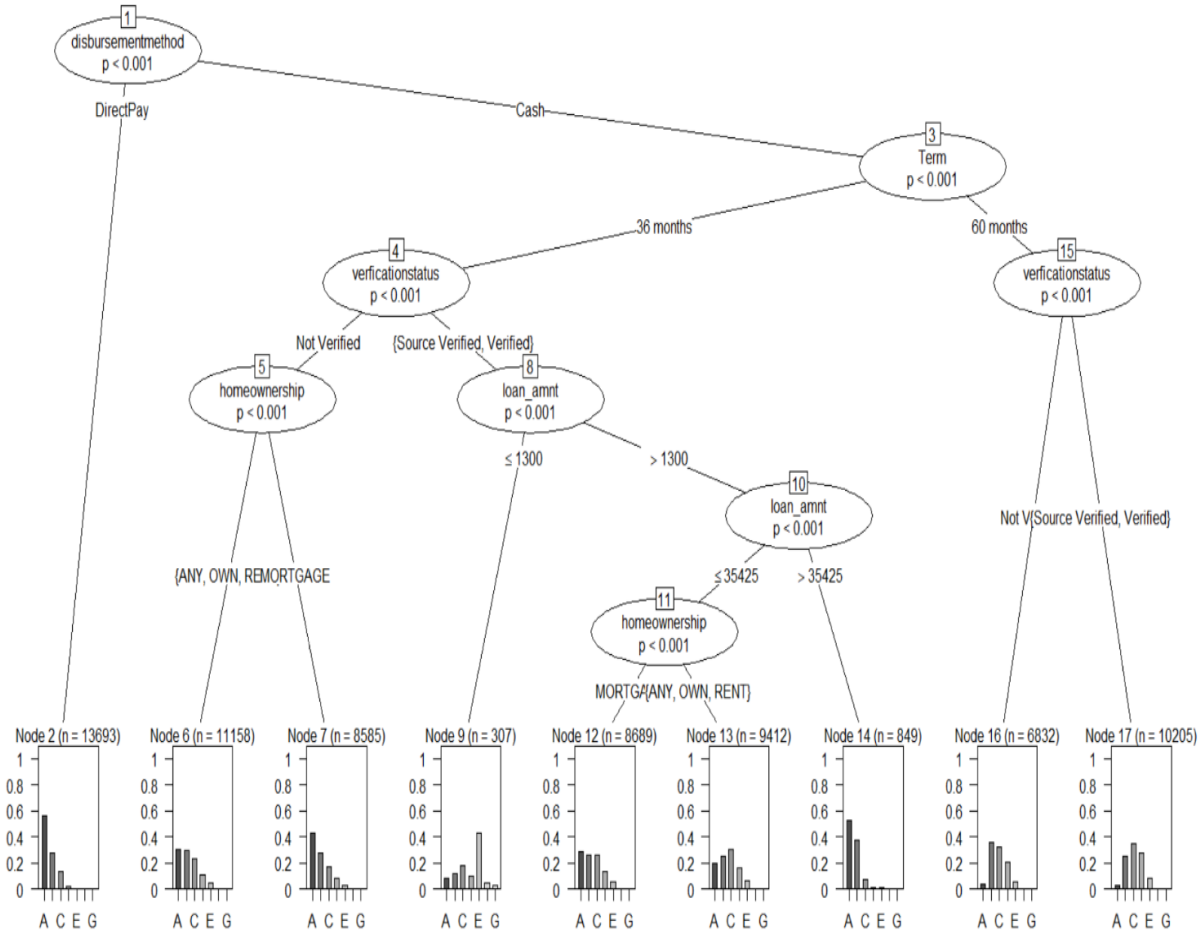

Figure 10: Decision Tree with Selected Independent Variables

This time, disbursement method is shown (Figure 10) as the most important variable, 'direct payments' belonging to grade 'A' with a probability of 60%. Term period of repayment is the next important variable with '60 months' of repayment period having a high probability of not falling into grade 'A'. When the term period is '36 months', verification status is considered with verified source of loan amount greater than $35,425 belonging to grade 'A'. Table 8 show the values of selected independent variables for grade 'A' and 'B' according to Decision Tree model.

Variable	Grade 'A'	Grade 'B'
Interest rate	< 8.81	< 12.96
Loan amount	35,425	35,425
Term	36 months	60 months
Verification status	Verified	Verified
Disbursement	Direct pay	Cash

Table 8: Summary of Loan Characteristics

After fitting the model into test data, it is seen that the accuracy of the model decreases drastically to 38.76% (Table 9). This shows that removing the independent variables causes the model to perform poorly with the same amount of training and test data.

testpred1	A	B	C	D	E	F	G
A	7660	5159	3636	1449	575	11	8
B	107	1052	953	626	165	10	1
C	913	2086	2842	1861	688	23	5
D	0	0	0	0	0	0	0
E	7	15	23	26	60	5	3
F	0	0	0	0	0	0	0
G	0	0	0	0	0	0	0

```
> 1-sum(diag(tab2))/sum(tab2)
[1] 0.6124662
```

Table 9: Confusion Matrix – Decision Tree with Selected Independent Variables

6 Discussion

The study focussed on identifying the factors that lead to high credit-risk in small business loans. Machine learning algorithms were used to learn from the data and predict the grade of the loans. It is seen that among the three models used, Decision Tree performs with the highest accuracy when all other independent variables are included in the data. Excluding some variables leads to poor performance as those are the factors that highly impact credit decisions. Random Forest also performs with an excellent accuracy of 99.83%, also showing the importance of each variable. Naïve Bayes performs with the lowest accuracy when compared to the other two models but greater than those achieved in previous studies (Table 10).

Model	Accuracy
Naïve Bayes	95.06%
Random Forest	99.83%
Decision Tree with all variables	99.99%
Decision Tree with selected variables	38.76%

Table 10: Models with their Accuracy

According to the models, for the given dataset, interest rate is the most important factor that decides the credit score of the loan i.e. grade of the loan. Interest rates of less than 8.81% are most likely to have the least credit-risk. This is followed by other factors such as the principal amount of REC loans taken, outstanding principal amount, disbursement method, term of repayment and loan amount. Banks can efficiently use such methods to set criteria on the loan applications by assigning values to each factor such that they would have the minimum credit-risk. SMEs can also decide on the quality of their loan applications when such values are made transparent to them and can work on improving their credit score.

However, the study suffers from some limitations such as; it does not take into account those loan applications that are applied with a joint account. The study is also limited to the information contained in the chosen dataset and does not take into account other external factors that could impact loan decisions. The Machine learning algorithms were chosen based on previous work, there may be other algorithms that could work well with a more complex data.

6 Conclusion and Future Work

This study aimed at identifying the most important factors that lead to high credit-risk in small business loans. It aimed at using machine learning techniques that could accurately predict the loan quality based on various attributes that are considered while evaluating loans. It is seen that Naïve Bayes, Decision Tree and Random Forest perform with good accuracy of more than 90% and speed on the given dataset containing 99, 699 rows and 25 variables. Interest rate is the primary factor that decides the credit risk of the loan and consequently its grade. Other factors such as term of repayment and disbursement method are also important factors that can be controlled by the SMEs before making a loan application. This would increase the chances of a loan being approved and would also minimize the risks incurred by the lending bank.

The proposed technique follows a transparent method that could increase financial knowledge among the SMEs and improve their relationship with banks. It would also significantly decrease loan decision time making the process efficient. However, factors such as the limited number of models and information used may be improved in future studies where a combination of other deep learning techniques can be implemented.

References

Angilella, S. and Mazzù, S. (2015) 'The financing of innovative SMEs: A multicriteria credit rating model', *European Journal of Operational Research*, 244(2), pp.540-554

Altman, E. and Saunders, A. (1998) 'Credit risk measurement: developments over the last 20 years', *Journal of Banking and Finance*, 21(11/12), pp. 1721-1742.

Belas, J., Smrcka, L., Gavurova, B. and Dvorsky, J. (2018) 'The Impact of Social and Economic Factorsin the Credit Risk Management of SME', *Technological and Economic Development of Economy*, 24(3), pp. 1215-1230.

Bengo, I. and Arena, M. (2019) 'The relationship between small and medium-sized social enterprises andbanks', *International Journal of Productivity and Performance Management*, 68(2), pp.389-406.

Briozzo, A., Vigier, H. and Martinez, L. (2016) 'Firm-Level Determinants of the Financing Decisions of Small and Medium Enterprises: Evidence from Argentina', *Latin American Business Review*, 17(3), pp. 245-268.

Davis, K., Maddock, R. and Foo, M. (2017) 'Catching up with Indonesia's fintech industry', *Law and Financial Markets Review*, 11(1), pp. 33-40.

Ferreira, F., Spahr, R., Gavancha, I. (2013) 'Readjusting trade-offs among criteria in internal ratings of credit-scoring: An empirical essay of risk analysis in mortgage loans', *Journal of Business Economics and Management*, 14(4), pp. 715-740.

Goncalves, T. et al., (2016) 'An Idiosyncratic Decision Support System for Credit Risk Analysis of Small and Medium-Sized Enterprises', *Technological and Economic Development of Economy*, 22(4), pp. 598-616.

Ju, Y. and Sohn, S. (2014). 'Updating a credit-scoring model based on new attributes without realization of actual data', *European Journal of Operational Research*, 234(1), pp.119-126.

Kelly, R., Brien, E. and Stuart, R. (2014) 'A long-run survival analysis of corporate liquidations in Ireland', *Small Business Economics*, 44(3), pp. 671-683.

Khandani, A., Kim, A. and Andrew, W. (2010) 'Consumer credit-risk models via machine-learning algorithms', *Journal of Banking and Finance*, pp 2767-2787.

Kljucnikov, A. and Belas, J. (2016) 'Approaches of Czech Entrepreneurs to Debt Financing and Management of Credit Risk', *Quarterly Journal of Economics and Economic Policy*, 11(2), pp. 343-365.

Lee et al., (2006) 'Mining the customer credit using classification and regression tree and multivariate adaptive regression splines', *Computational Statistics and Data Analysis*, 50, pp. 1113-1130.

Monika, S. (2016) 'The Relationship between the Risk of a Change of the Interest Rate and the Age of Entrepreneurs among Slovak SMEs', *Journal of Competitiveness*, 8(3), pp. 125-138.

Navaretti, G., Calzolari, G. and Pozzolo, A. (2015) 'How SME funding risk is allocated' | VOX, CEPR Policy Portal. [online] Voxeu.org. Available at: https://voxeu.org/article/how-sme-funding-risk-allocate [Accessed 20 Mar. 2019].

Pandey et al., (2017) 'Credit risk analysis using machine learning classifiers', *2017 International Conference on Energy, Communication, Data Analytics and Soft Computing*, p. 1850.

Sun, H. and Guo, M. (2015) 'Credit risk assessment model of small and medium-sized enterprise based on logistic regression', *2015 IEEE International Conference on Industrial Engineering and Engineering Management*, pp. 1714.

Zhu, Y., Xie, C., Wang, G. and Yan, X. (2016a) 'Predicting China's SME Credit Risk in Supply Chain Finance Based on Machine Learning Methods', *Entropy*, 18(5), p. 195.

Zhu, Y., Xie, C., Wang, G. and Yan, X. (2016b) 'Comparison of individual, ensemble and integrated Ensemble machine learning methods to predict China's SME credit risk in supply chain finance', *Neural Computing and Applications*, 28(1), pp. 41-50.