FACULTY OF COMPUTING AND INFORMATICS

SEMESTER 2-2020/2021

KD14403 FUNDAMENTAL OF DATA SCIENCE

GROUP ASSIGNMENT

LOAN APPROVAL USING MACHINE LEARNING

GROUP MEMBERS:

No	Name	Matric Number	Role
1	LEE YI FENG	BI20110003	Conclusion & Future Work, Compiler, Build decision tree model, Result & Discussion
2	CHAN KAI KHEE	BI20110118	Trends & Challenges, Build logistic regression model, Result & Discussion
3	CHIN KAH KHING	BI20110081	Abstract & Introduction, Build logistic regression model, Result & Discussion
4	FOO RUI ZHE	BI20110088	Methodology, Build random forest model, Result & Discussion
5	LEE BING QIAN	BI20110252	Problem Statement, Build random forest model, Result & Discussion
6	NGU YUN LIN	BI20110144	Opportunities, Build decision tree model, Result & Discussion

CONTENTS

ABSTRACT

In the contemporary day and age, taking loans from banks and other financial organizations has become quite frequent. As the occurrences of non-payment circumstances, prediction of the possibility of payment loan by the clients is enormous for the lenders to limit the occurrences and protect themselves from any losses. In this paper, machine learning (ML) techniques are used to predict loan approval as ML is quite beneficial when it comes to forecasting outcomes for massive amounts of data. Three models were built in this paper which are Decision Tree (DT), Random Forest (RF) and Logistic Regression (LR). The loan dataset was divided randomly into train and test dataset whereas the train dataset is trained to learn an effective mapping of inputs to output variables and the test dataset is used to make predictions as well as validate the models. After the validation of the test dataset, accuracy for every model was measured by using a confusion matrix. The experimental results conclude that the accuracy of Random Forest is the highest. It is the most suitable model to predict the loan approval among all of the models used in this paper.

Keywords- Loan, Machine Learning, Train, Test, Prediction.

INTRODUCTION

The term loan is a form of debt incurred by an individual or other entity. Money is given to another party from the lender in exchange for repayment of the loan principal amount plus interest or any finance charges (Julia Kagan, 2021) The lender could be a corporation, financial institution, or government. The loan terms must be acknowledged and agreed by each party before any money is advanced. In the loan process, there are some details needed to be provided to the lender such as the reason for the loan, financial history, dependents etc. The lender will review and analyze the information that was provided before proceeding to the loan. They need to ensure that the borrower has the ability to repay the loan in a certain period of time.

Nowadays, there is a lot of clients that fail to do the repayment (M. A. Sheikh, A. K. Goel and T. Kumar, 2020) within the scheduled time and some clients even run away from the debts. As a result of this phenomenon, the lender encounters those unwanted financial burdens while the applicants also incur heavier losses as they have to let go of their properties or assets to facilitate payment. To limit the losses suffered from the lender, the loan approval prediction system acts as a crucial process. This system allows the lender to forecast whether the applicant is trustworthy and verify the ability of the applicant to repay the loan by analyzing the information provided by the borrowers.

Loan approval prediction is pretty helpful for the lenders and the applicants as well because it also allows the applicants to grasp their financial ability and may decrease their burden if they overestimate their ability to pay the loan. The demand for loans is increasing day by day, so the lender has to find the best way to analyze and verify the right deserving loan applicant (Kumar Arun, Garg Ishan and Kaur Sanmeet, 2016). Having a trustworthy borrower also smoothens the operation and benefits the organizations. The purpose of this paper is to come up with a way which is more efficient, quick, immediate and easy to carry out the selection of the qualified applicants. Once the model in the loan prediction system learned the weight of each feature taking into account loan processing and on new test data the same features are processed with respect to their associated weight (J.Tejaswini, T.Mohana Kavya, R. Devi Naga Ramya, P.Sani Triveni and Venkata Rao Maddumala, 2020). The loan prediction system is able to compute the data of the features automatically and predict the ability for an applicant to get the loan.

PROBLEM STATEMENT

With significant advancements in big data and machine learning techniques in recent years, a range of financial industries have used machine learning approaches in financial management. Previous research indicates that implementing machine learning algorithms could help the banking sector handle financial risks more efficiently (Mitchell, 1999). A loan approval prediction system built with the models that anticipates the consumer loan approval is created and adopted by the banking sectors. In this system, Logistic Regression is typically used to forecast if a consumer is qualified for a loan application (Tong, Mues, and Thomas, 2012).

However, according to some other research, there are other ways that might function more efficiently than the standard machine learning methodology, Logistic Regression, in loan prediction (Granström, and Abrahamsson, 2019). Several other models have been shown to be helpful in projecting loan approval in recent years. As a result, the issue raised is, which model performs the best in loan approval prediction?

Hence, the focus of our research will be on determining which model, among Decision Tree, Random Forest, and Logistic Regression, performs the best with the highest accuracy in loan approval prediction.

RESEARCH QUESTIONS

1. What kind of applicant qualified for loan application?
2. Is the applicant eligible for loan application?

RESEARCH OBJECTIVES

1. To investigate the Supervised Learning Models to predict the loan approval.
2. To evaluate and choose the most suitable model to select the applicant which can successfully get the loan.

STATE OF ART / LITERATURE REVIEW

Loan is important as it is the core business part for almost all of the banks because the main bank's assets directly come from the profit earned from the loan distribution. Bank or financial companies loan approval mostly has been made after a regress process of verification and validation. However, there are still misgivings about whether the selected applicants were deserving of the right applicant. Therefore, Loan Approval Prediction is made to predict the applicant in a safe way and used machine learning techniques during the validation process. Loan Approval Prediction system can calculate the weight of each feature in loan processing automatically. New test data are also tested on the same features with respect to their associated weight. The dataset was collected and featured by using info gain of features. The model is then trained and tested on the training dataset and testing dataset respectively (Arun, K., Sanmeet, K., and Ishan, G. 2014). Furthermore, there are also 6 machine learning methods used in their research which are decision trees, random forest, Support Vector Machine (SVM), linear models, neural network and adaboost. The parameters were set for each machine learning model. For the results, it is safely concluded that the product is a highly efficient component as it worked properly and met all Banker requirements.

The technical world is advancing toward complete automation to encounter any problems in a speed and accurate way. However, decision taking is attained by probabilistic and predictive approaches developed by various machine learning algorithms. Machine Learning is a tool which facilitates development of analytical models without explicit programming. Various machine learning algorithms are developed to tailor to the problem requirements. All the leading-edge industries are now utilizing the capabilities of machine learning to gain higher sales growth and statistics have shown that they are getting positive results. With institutions generating more and more data, exploitation of data manually becomes difficult, hence machine learning, having the capability of analytical modelling is sought to, as a solution. This paper adheres to logistic regression as a machine learning tool in order to actualize the predictive and probabilistic approaches to a given problem of loan approval prediction. Many models were henceforth developed to account for dichotomous behaviour of the outcome variable. The Logistic Regression model was chosen over the other models because of its mathematical clarity and flexibility. Once the data is properly rectified and idealized, it is used to fit into the model and ready to train the model. To minimise the overfitting of the data to train the model, few features are selected to train the model which are 'Credit history', 'self-employed', 'property area', 'education'. If the probability of loan getting approved is more than 0.5 then, the predicted value for the set of attributes will be that the loan will get approved and vice versa. (Vaidya, A.2017)

Peer-to-peer (P2P) lending has become a significant trend in the lending industry in recent years, thanks to the advancement of big data and online finance. P2P lending is an online platform that connects lenders and investors to build credit connections and complete transaction procedures. However, as demand for this platform grows, the main concerns emerge: financial risks such as liquidity risks, as well as legal risks caused by ineffective Internet finance laws and regulations. Therefore, machine learning techniques are now widely used in the online financial industry to manage financial risk. A research conducted by the researchers to study loan default predictions with Random Forest Algorithm. In the study, four models were created with different algorithms including Random Forest, Decision Tree, Support Vector Machine (SVM) and Logistic Regression. Besides, a dataset collected from Lending Club including 115,000 loan data of individuals and 102 variables was employed to train the models. Based on the results of the study, the Random Forest classifier predicts loan default more accurately than other classifiers built with different machine learning algorithms. It is the most accurate, with a 98 percent accuracy rate which is more accurate than the Decision Tree which only has 95 percent accuracy rate. (Zhu et al., 2019)

In this research, they found that financial institutions have a large amount of data on their borrowers, which can be used to predict the probability of borrowers defaulting their loan or not. The purpose of this research was to analyse individual loan defaults in Kenya using the logistic regression model. This study employed a quantitative research design, it deals with individual loans defaults as group characteristics of a borrower. The data was pre-processed by seeding using R- Software and then split into training dataset and test data set. The train data was used to train the logistic regression model by employing a Supervised machine learning approach. The test data set was used to do cross-validation of the developed logistic model which later was used for analysis prediction of individual loan defaults. The logistic regression model predicted 303 defaults from the train data set, 122 non-defaults and misclassified loans were 56 and 69. The model had an accuracy of 0.7727 with the train data and

0.7333 with the test data. The logistic regression model showed a precision of 0.8440 and 0.8244 with the train and test data respectively. The study recommended the use of logistic regression in conjunction with supervised machine learning approach in loan default prediction in financial institutions. In this study, a logistic model was used for the analysis of individual loan defaults. This study was motivated by the increasing need to explain how individual loan defaults relates to different variables of interest in the Kenyan financial institutions as well as determine how to mitigate the menace of loan defaults. (Mong'are et al., 2019).

Data mining is the process of examining data from many angles and extracting meaningful information from it. It is at the heart of the knowledge-creation process. The several stages that go into extracting knowledge from raw data. Classification, clustering, association rule mining, prediction and sequential patterns, neural networks, regression, and other data mining techniques are different data mining techniques. The most widely used data mining approach is classification, which uses a group of pre-classified samples to create a model that can categorise the whole population of information. The classification method is particularly well suited to fraud detection and credit risk applications. Decision Tree-based Classification Algorithms are widely used in this method. A training set is used to develop a model as a classifier that can categorise data objects into their respective classes in classification. The model is validated using a test set. Due to the massive growth of data in the banking industry, data mining techniques can be used widely to analyse and predict the trends. This will be very effective in certain areas such as marketing, risk management and customer relationship management. A loan credibility prediction system is developed in this research to assist companies in making the best choice on whether to approve or reject client loan requests. This would undoubtedly aid the banking industry in the establishment of effective distribution channels. The prediction is based on the Decision Tree Induction Algorithm. Other techniques that outperform current data mining methods must be incorporated and tested for the domain. (Sivasree M.S. and Rekha Sunny T., 2015).

Loan approval is a critical process for banking institutions. Loan Prediction is very helpful for employee of banks as well as for the applicant also. A quick, immediate and easy way to choose the deserving applicants is provided by applying different machine learning. In this paper, three machine learning algorithms, Logistic Regression (LR), Decision Tree (DT) and Random Forest (RF) are applied to predict the loan approval of customers. The Loan Prediction System allows you to jump to a specific application and check it on a priority basis. The entire prediction process is carried out in private, and no stakeholders are able to influence the outcome. The Loan Prediction System can automatically calculate the weight of each feature involved in loan processing, and the same features are processed in relation to their associated weight on new test data. The training data set is now provided to the machine learning model, and the model is trained using this data set. Every new applicant's information entered on the application form serves as a test data set. Following the testing operation, the model predicts whether the new applicant is a fit case for loan approval or not based on the inferences it draws from the training data sets. (Tejaswini et al., 2020)

METHODOLOGY

Dataset

Secondary data was found on a website called Kaggle and used in the research. The dataset contains 614 observation and 13 variables.

Variables	Data Type
Loan ID	Character
Gender	Character
Married	Character
Dependents	Integer
Education	Character
Self_Employed	Character
ApplicantIncome	Integer
CoapplicantIncome	Integer
LoanAmount	Integer
Loan_Amount_term	Integer
Credit History	Integer
Property_Area	Character
Loan_Status	Character

The variables in the dataset act as the features to predict the loan approval of the applicant during the validation process. There were total of 149 missing values found in the dataset. The variables which contained missing values were Credit History (50), Self-employee (32), loan amount (22), dependents (15), loan amount term (14), gender (13) and married (3). The missing values were fixed during data-preprocessing phase.

Techniques

The project was conducted by using machine learning techniques. Machine learning is a concept that provides machine to learn from the real-world interaction and observation and has the ability to behave like humans and their performance can be improved by data given as input (Maddan,2021). Banking sector and finance are the popular example of Machine Learning application. Machine learning can be divided into 3 steps which are file execution, forecasting models and prediction which are also used to construct the flow of the project. Observation process and conclusion making were conducted with machine learning models. The models that were used in this project were decision tree, random forest and logistic regression. All of the methods used in the project were under supervised learning methods.

Models

The reason for using decision tree model, random forest and logistic regression in this research was to perform classification, regression and prediction. Decision tree is one of the most popular algorithms used for classification and regression. It consists of several branches, leaf nodes and root nodes. A tree-like structure is generated by classifying the instance and utilizing a Recursive Partitioning Algorithm (RPA). Furthermore, Random Forest model was chosen because it has the similar function with Decision Tree but with the extra advantages of immunity to overfitting, more efficient on large databases and has an accurate classification and regression. Lastly, logistic regression was used for prediction.

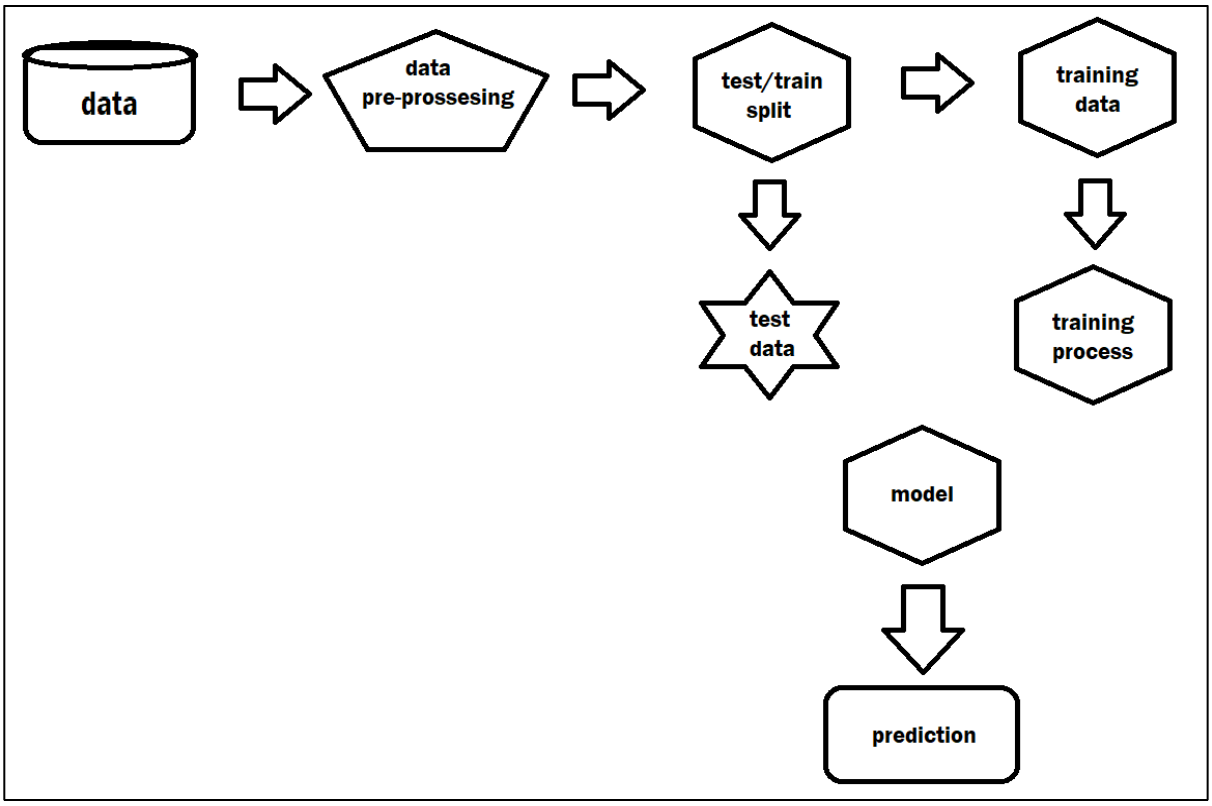


Figure : Schematic model construction and evaluation

Data pre-processing

The real-world data is mostly incomplete where it contains some missing values, noises, errors and outliers. Therefore, data preprocessing is important in order to prepare a new complete data to make it more suitable before modelling (Pandey, 2020). It also improved the efficiency and accuracy of a machine learning model. In the data pre-processing step, the data was first loaded into the R program by using read() function. The data was then analyzed by using head() function and summary() function. This was to identify the missing values or faulty data in the dataset. Furthermore, miss_var_summary() from nianar library was used to summarise each variable and calculate the sum of missing data in each variable and its missing percentage. The Aggr() function from the VIM library was used to visualize and analyze the missing value by using graphical method. Then, the missing data was imputed by mice() function (Alice, 2018). Mice (Multivariate Imputation via Chained Equations) assumes the missing data are Missing at Random (MAR) and imputes the data on a variable-by-variable basis by specifying an imputation model per variable. Moreover, complete() function was used to combine the data set that has been processed by mice() with the original dataset. Miss_var_summary() function and missmap() function were used to verify any missing data value in the clean dataset. The new clean data set was analyzed by using summary() function. The features in the dataset were also categorized into categorical and numerical variables.

Data Visualization

Graphical method was used to visualise the data set. Graphs such as histogram, bar charts were used during data visualization. Data visualization is important as it gives us the information of the data through maps or graphs thus helps us to understand more of our data (Brush & Burns, 2020). The trends, patterns and outliers within large data sets are easier to be identified. Histogram was used to determine the outliers and the distribution in each numeric variable while grouped bar charts were used to determine the correlation between each categorical variable with the loan status.

Modelling

Before modelling, the applicant income and co-applicant income were added together to form total income for better classification. The log() function was used in some numerical features to obtain normal distribution of the data. The data was then split into training data and test data. Training data was used in model evaluation for quantifying the performance of a model while test data was used in model validation to make sure the model built will work in production.

Decision Tree

The Decision Tree model was chosen as one of the predictive models in this research. Decision tree learning takes into the observations about an item and predicts the item's value. In this method, the decision tree is also known as the classification tree. The nodes in the tree represent data rather than decisions. Each branch contains classification rules and are associated with a particular class label. The rules can be expressed as if-then clauses, where each data value or decision forms a clause. The decision tree model was first trained with all of the variables. The significant variables were then determined and used on the second model. The second model was pruned in order to improve the model accuracy.

Logistic Regression

In Logistic regression model, significant variables were examined before training the model. Then, 2 models were built to determine which model would perform the best in predicting loan approval. The AIC score was determined to examine the quality of the model. To prevent overfitting, the more important variables were filtered out according to Pr(|z|) value. Hence, the second model was built with the chosen significant variables to enhance the model quality. After that, the model was used to make predictions on test data to evaluate the model performance. Confusion matrix was used to calculate the accuracy of the prediction. Furthermore, ROC graph also used to calculate the accuracy of the model.

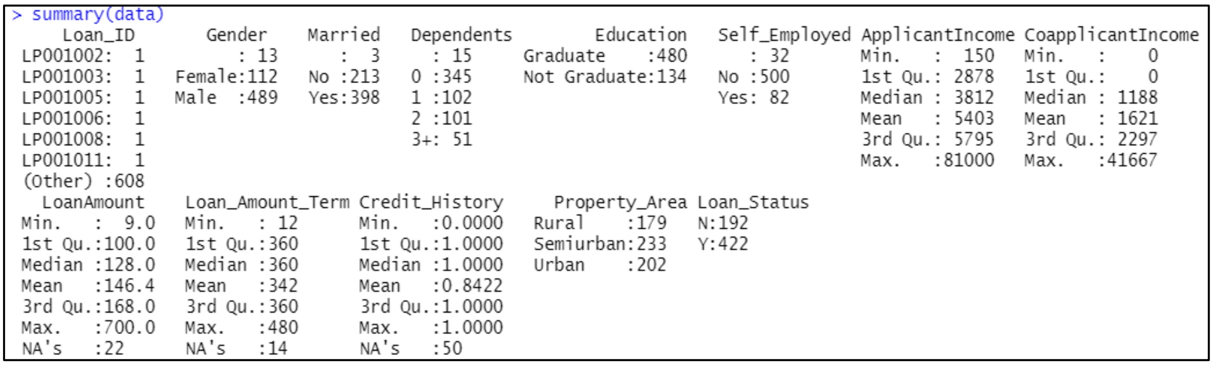
Random Forest

Random forest was used for classification and regression. It constructs multiple decision trees at training time to get more stable and accurate predictions. The trees were set to 500 which is the default number of trees in the random forest model. The model was first constructed by predicting the loan status with all of the variables except loan id, total income, applicant income and co applicant income. Then, VarImplot() was used to determine which variables are more important. Hence, the random forest model was built with the more significant variables to increase the accuracy of the model.

The accuracy of the models were then compared to find the most suitable model for the project.

Discussion

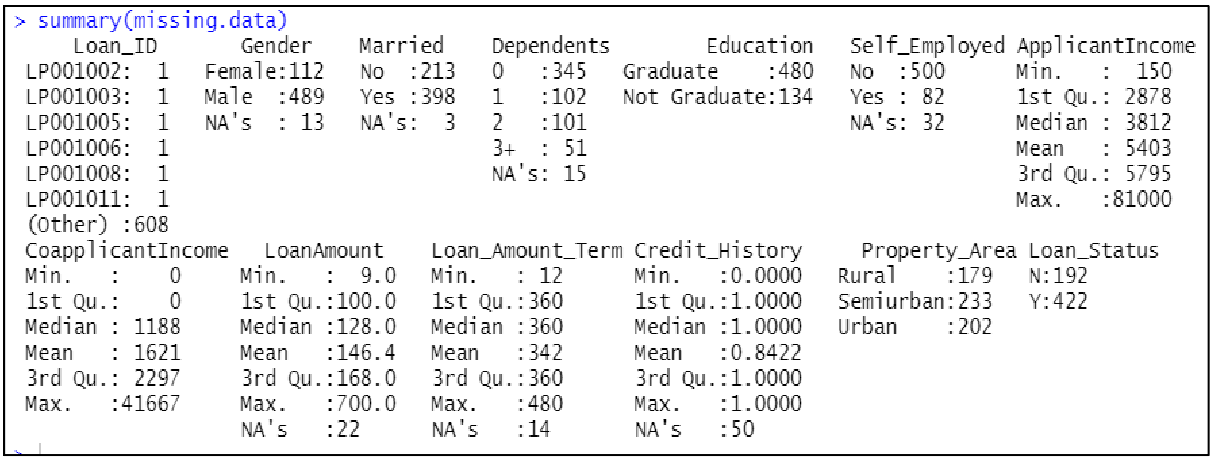
The data set was analyzed after loaded in R program by using head() and summary() function.



```
> summary(data)
     Loan_ID       Gender     Married    Dependents        Education    Self_Employed ApplicantIncome CoapplicantIncome
 LP001002:  1         : 13       :  3       : 15     Graduate    :480      : 32     Min.   :  150   Min.   :    0    
 LP001003:  1   Female:112   No :213     0 :345     Not Graduate:134   No :500     1st Qu.: 2878   1st Qu.:    0    
 LP001005:  1   Male  :489   Yes:398     1 :102                        Yes: 82     Median : 3812   Median : 1188    
 LP001006:  1                            2 :101                                    Mean   : 5403   Mean   : 1621    
 LP001008:  1                            3+: 51                                    3rd Qu.: 5795   3rd Qu.: 2297    
 LP001011:  1                                                                      Max.   :81000   Max.   :41667    
 (Other) :608                                                                                                       
   LoanAmount    Loan_Amount_Term Credit_History     Property_Area Loan_Status
 Min.   :  9.0   Min.   : 12      Min.   :0.0000   Rural    :179   N:192      
 1st Qu.:100.0   1st Qu.:360      1st Qu.:1.0000   Semiurban:233   Y:422      
 Median :128.0   Median :360      Median :1.0000   Urban    :202              
 Mean   :146.4   Mean   :342      Mean   :0.8422                              
 3rd Qu.:168.0   3rd Qu.:360      3rd Qu.:1.0000                              
 Max.   :700.0   Max.   :480      Max.   :1.0000                              
 NA's   :22      NA's   :14       NA's   :50                                  
```

Figure : Summary of data

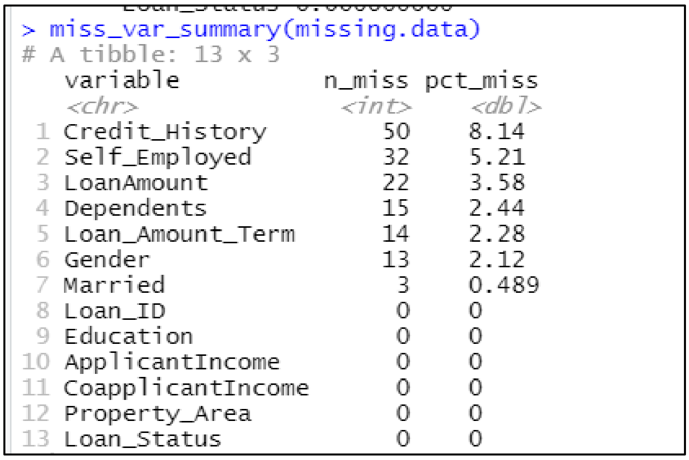
The summary of data showed that there were a few missing values and some NA strings. In oder to determine all of the NA's values, the data renamed into missing data and the NA's strings were replaced by the word "NA". The missing data were then summarized using summary function.



```
> summary(missing.data)
     Loan_ID       Gender      Married     Dependents        Education    Self_Employed ApplicantIncome
 LP001002:  1   Female:112   No  :213   0   :345   Graduate    :480   No  :500     Min.   :  150  
 LP001003:  1   Male  :489   Yes :398   1   :102   Not Graduate:134   Yes : 82     1st Qu.: 2878  
 LP001005:  1   NA's  : 13   NA's:  3   2   :101                      NA's: 32     Median : 3812  
 LP001006:  1                           3+  : 51                                   Mean   : 5403  
 LP001008:  1                           NA's: 15                                   3rd Qu.: 5795  
 LP001011:  1                                                                      Max.   :81000  
 (Other) :608                                                                                     
 CoapplicantIncome   LoanAmount    Loan_Amount_Term Credit_History     Property_Area Loan_Status
 Min.   :    0     Min.   :  9.0   Min.   : 12      Min.   :0.0000   Rural    :179   N:192      
 1st Qu.:    0     1st Qu.:100.0   1st Qu.:360      1st Qu.:1.0000   Semiurban:233   Y:422      
 Median : 1188     Median :128.0   Median :360      Median :1.0000   Urban    :202              
 Mean   : 1621     Mean   :146.4   Mean   :342      Mean   :0.8422                              
 3rd Qu.: 2297     3rd Qu.:168.0   3rd Qu.:360      3rd Qu.:1.0000                              
 Max.   :41667     Max.   :700.0   Max.   :480      Max.   :1.0000                              
                   NA's   :22      NA's   :14       NA's   :50                                  
```

Figure 2 : summary of missing data

The summary of the missing data set showed there were some missing values in gender, married, dependents, self-employed, loan amount, loan amount term and credit history variables. The missing data values were then analyzed by using miss_var_summary().



```
> miss_var_summary(missing.data)
# A tibble: 13 x 3
   variable          n_miss pct_miss
   <chr>              <int>    <dbl>
 1 Credit_History        50    8.14
 2 Self_Employed         32    5.21
 3 LoanAmount            22    3.58
 4 Dependents            15    2.44
 5 Loan_Amount_Term      14    2.28
 6 Gender                13    2.12
 7 Married                3    0.489
 8 Loan_ID                0    0
 9 Education              0    0
10 ApplicantIncome        0    0
11 CoapplicantIncome      0    0
12 Property_Area          0    0
13 Loan_Status            0    0
```

Figure 3 : Summary of each missing variables

Miss_var_summary() function summarized that there were 50 missing values in credit history variable, 32 in self employed variable, 22 in loan amount variable, 15 in dependents variable, 14 in loan amount term variables, 13 in gender variables and 3 in married variables which hold 8.14, 5.21, 3.58, 2.44, 2.28, 2.12 and 0.489 percentage of missing respectively. In order to understand more about the missing value, Aggr() function from the VIM library was used to visualize the missing data in the dataset.

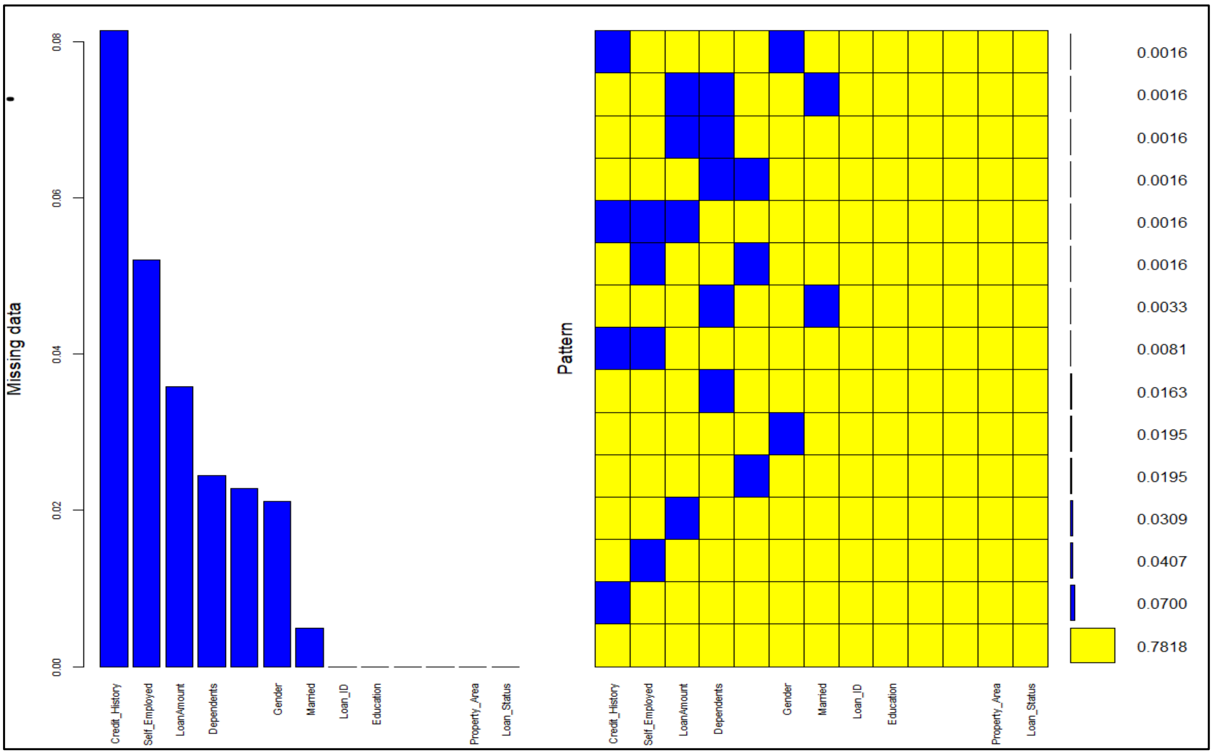
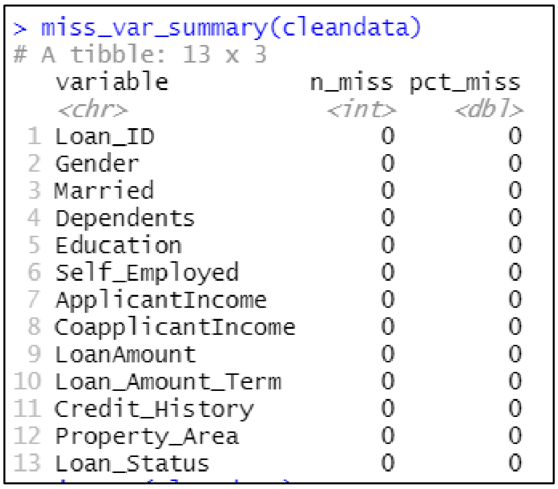


Figure 4 : Missing values were visualized by Aggr() function

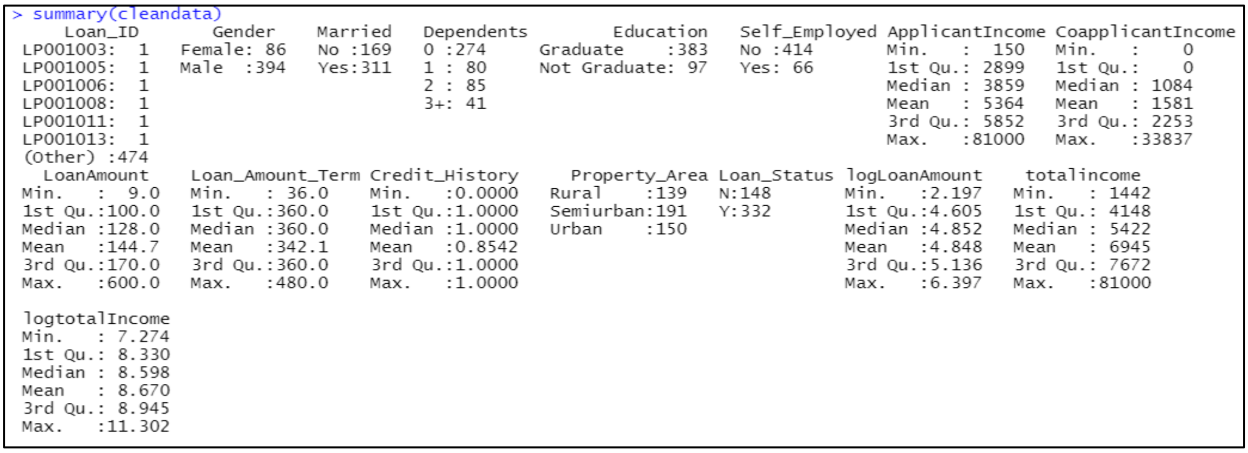
Then, the missing values were handled using mice() function instead of deleting it due to the small size of dataset. The CART method which stands for classification and regression trees was used and the number of iterations for each imputation was set to 2 in mice function. After that, the second round of imputed data was chosen and merged with the original data using complete() function. The merged data was labeled as clean data. Then, the clean data was checked by using miss_var_summary to make sure there were no more missing values.



```
> miss_var_summary(cleandata)
# A tibble: 13 x 3
   variable           n_miss pct_miss
   <chr>               <int>    <dbl>
 1 Loan_ID                 0        0
 2 Gender                  0        0
 3 Married                 0        0
 4 Dependents              0        0
 5 Education               0        0
 6 Self_Employed           0        0
 7 ApplicantIncome         0        0
 8 CoapplicantIncome       0        0
 9 LoanAmount              0        0
10 Loan_Amount_Term        0        0
11 Credit_History          0        0
12 Property_Area           0        0
13 Loan_Status             0        0
```

Figure: summary of missing data in processed dataset

After the missing value were imputed, the clean data was summarized again.



```
> summary(cleandata)
      Loan_ID        Gender      Married    Dependents        Education    Self_Employed ApplicantIncome CoapplicantIncome
 LP001003:  1   Female: 86   No :169   0 :274      Graduate    :383   No :414       Min.   :  150   Min.   :    0    
 LP001005:  1   Male  :394   Yes:311   1 : 80      Not Graduate: 97   Yes: 66       1st Qu.: 2899   1st Qu.:    0    
 LP001006:  1                          2 : 85                                       Median : 3859   Median : 1084    
 LP001008:  1                          3+: 41                                       Mean   : 5364   Mean   : 1581    
 LP001011:  1                                                                       3rd Qu.: 5852   3rd Qu.: 2253    
 LP001013:  1                                                                       Max.   :81000   Max.   :33837    
 (Other) :474                                                                                                        
   LoanAmount     Loan_Amount_Term Credit_History    Property_Area Loan_Status logLoanAmount    totalincome   
 Min.   :  9.0   Min.   : 36.0    Min.   :0.0000   Rural    :139   N:148       Min.   :2.197   Min.   : 1442  
 1st Qu.:100.0   1st Qu.:360.0    1st Qu.:1.0000   Semiurban:191   Y:332       1st Qu.:4.605   1st Qu.: 4148  
 Median :128.0   Median :360.0    Median :1.0000   Urban    :150               Median :4.852   Median : 5422  
 Mean   :144.7   Mean   :342.1    Mean   :0.8542                               Mean   :4.848   Mean   : 6945  
 3rd Qu.:170.0   3rd Qu.:360.0    3rd Qu.:1.0000                               3rd Qu.:5.136   3rd Qu.: 7672  
 Max.   :600.0   Max.   :480.0    Max.   :1.0000                               Max.   :6.397   Max.   :81000  
                                                                                                              
 logtotalIncome  
 Min.   : 7.274  
 1st Qu.: 8.330  
 Median : 8.598  
 Mean   : 8.670  
 3rd Qu.: 8.945  
 Max.   :11.302  
```

Figure : Summary of processed data

The summary of the data also showed that there were missing value or empty variables left. Then, All the variables were analyzed again. Data visualization step were performed after making sure there were no missing value left. The numerical variables such as loan amount, applicant income, co-applicant income, loan amount term and credit history were the first to be visualize.

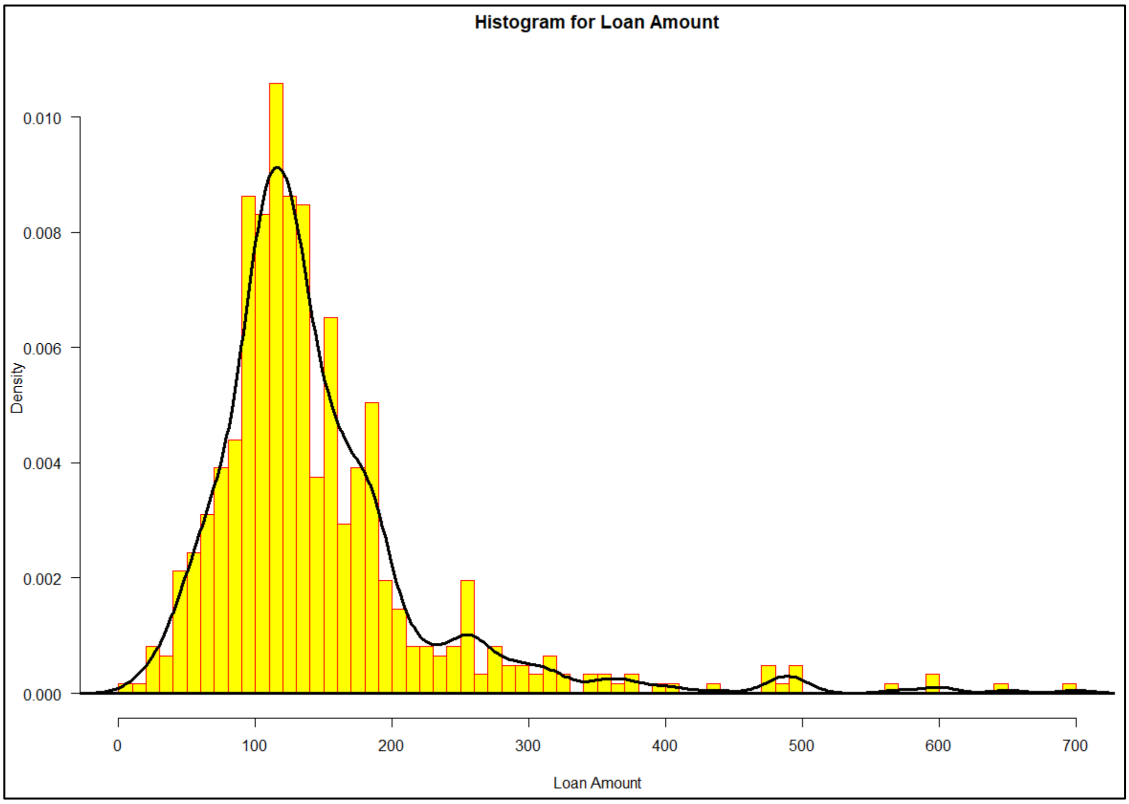


Figure 6: Histogram for Loan Amount

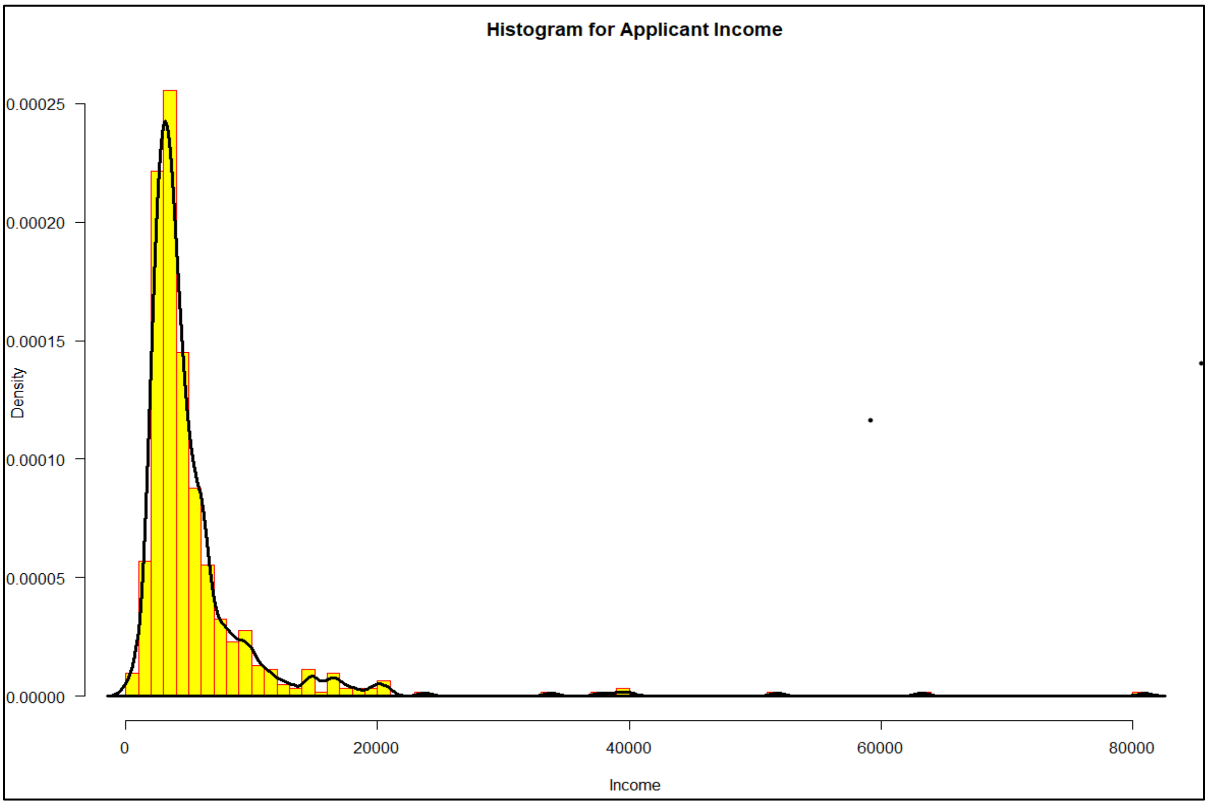


Figure 7: Histogram for applicant income

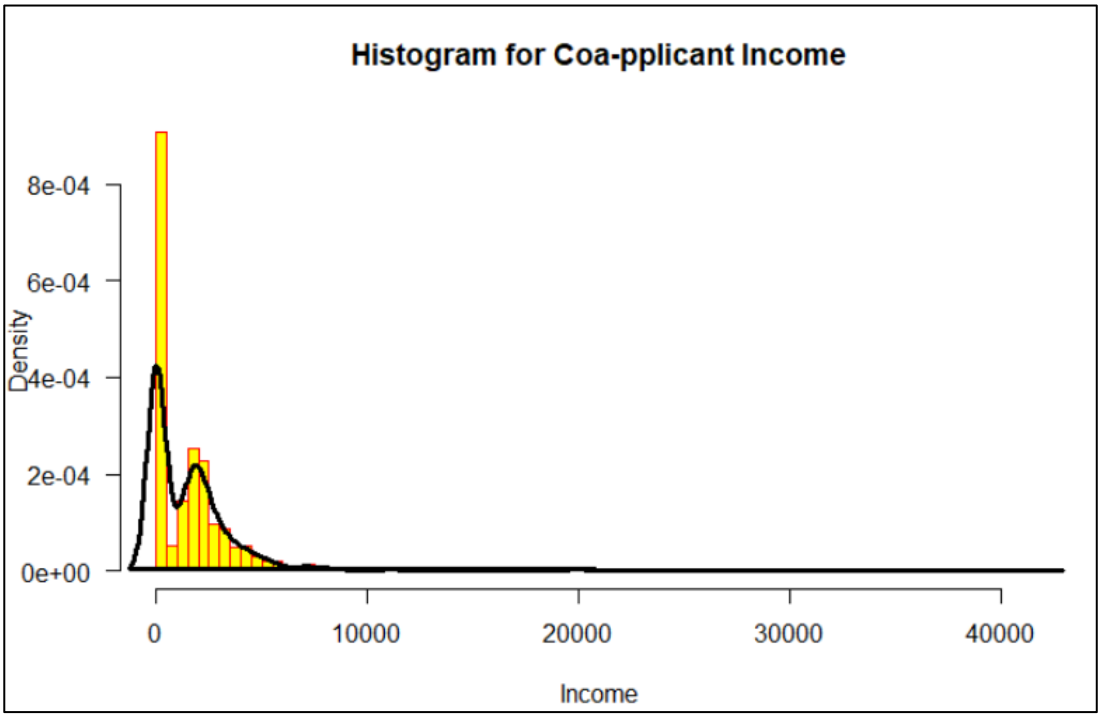


Figure 8 : Histogram for Co-applicant Income

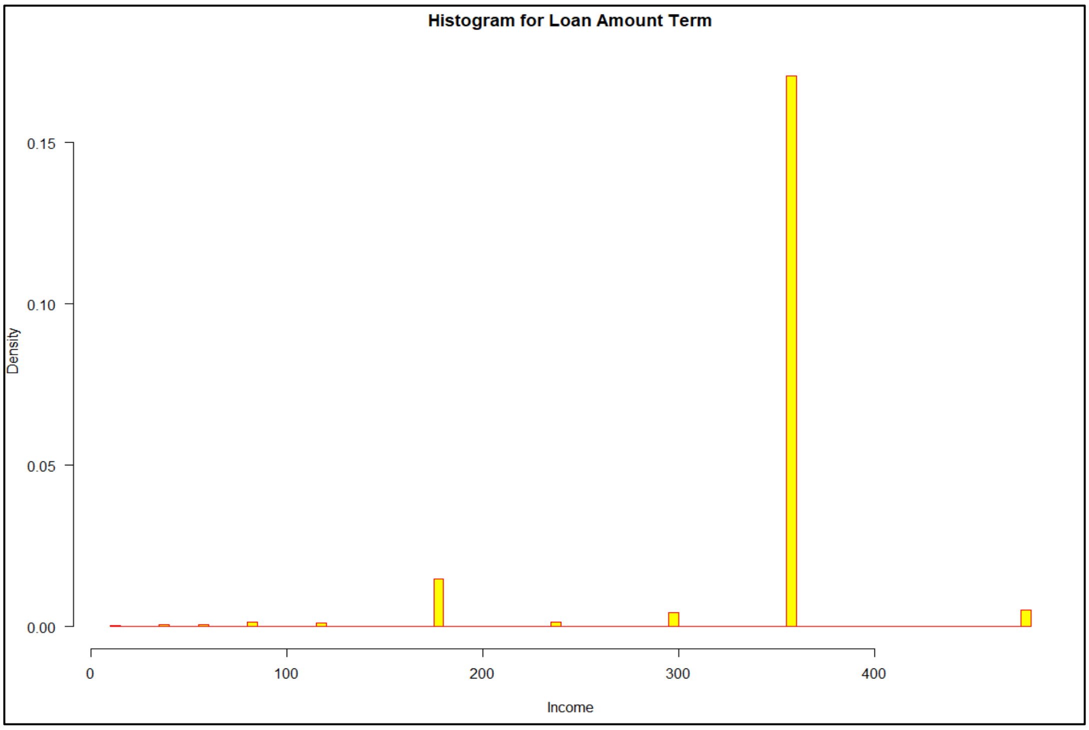


Figure 9 : Histogram for Loan Amount Term

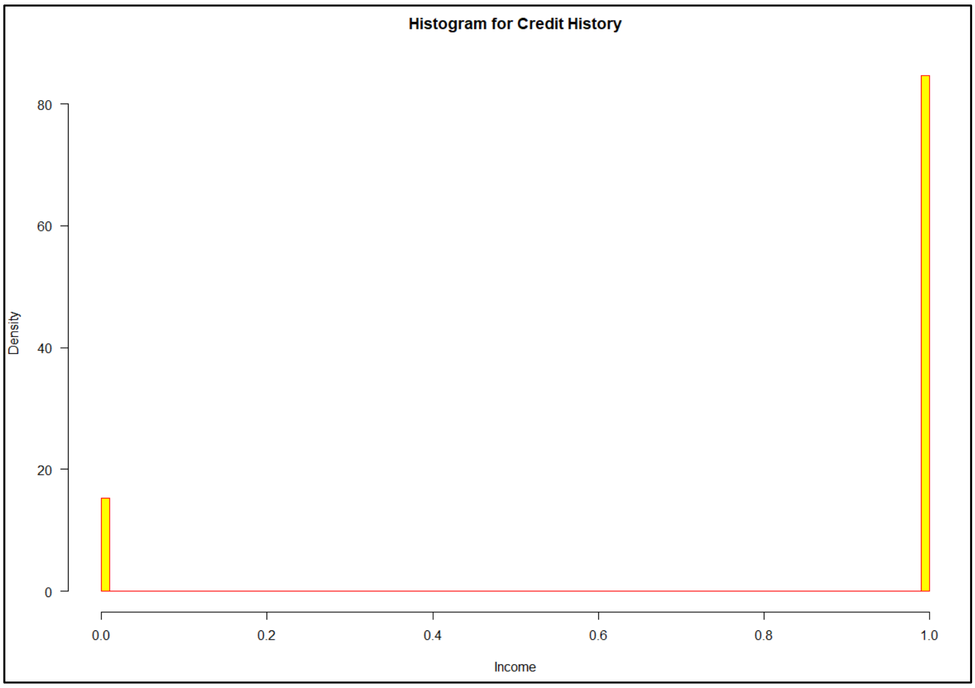


Figure 10 : Histogram for Credit History

The histograms of loan amount, applicant income and co-applicant income were right skewed which showed that there were extreme values found in loan amount and applicant income variable. Before dealing with extreme values, the applicant income and co-applicant were added together to form total income as it is under in the same category. To achieve a normal distribution value, log() function was used on both numerical variables.

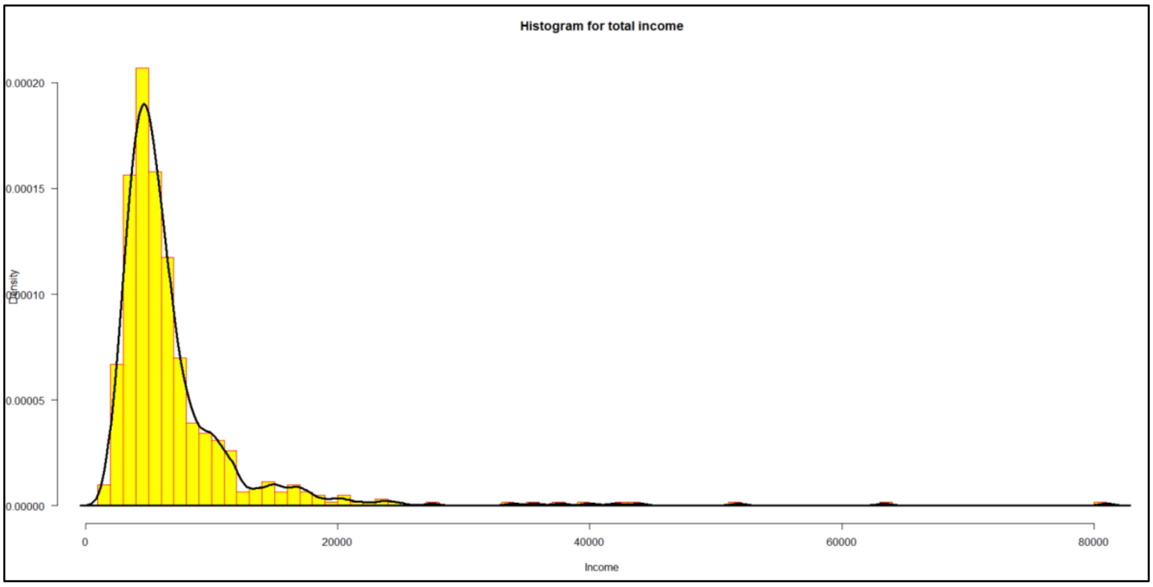


Figure : Histogram for total income

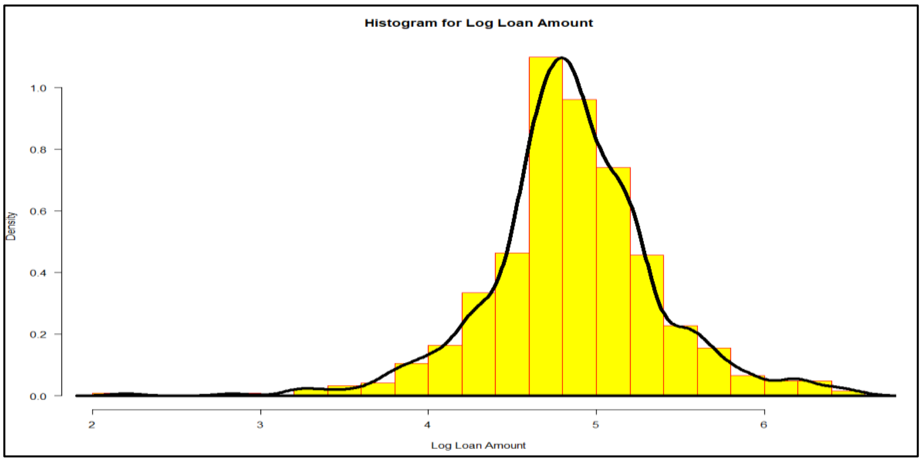


Figure 11 : Histogram of Log Loan Amount

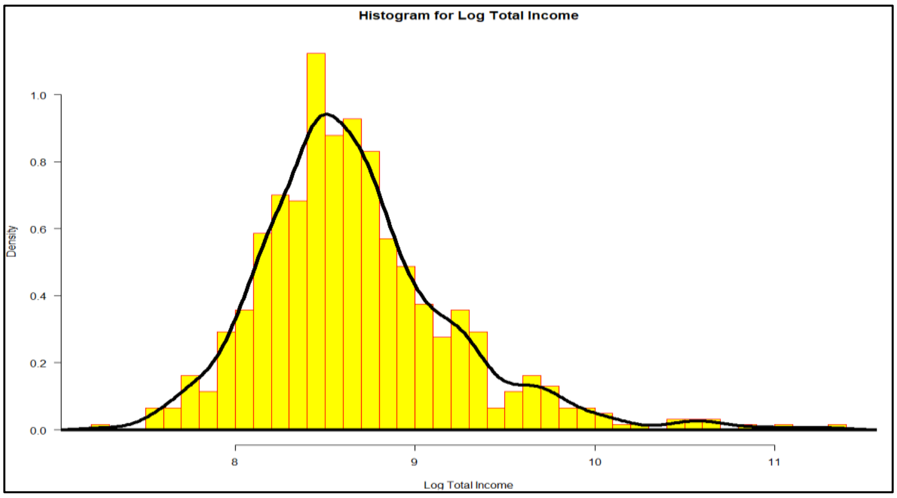


Figure 12 : Histogram of Log Total Income

The loan amount and total income were less skewed which were nearly normal distributed after log function was applied. Then, the categorial variables were visualized by using bar charts to find the correlation between each of the categorial variables with loan status.

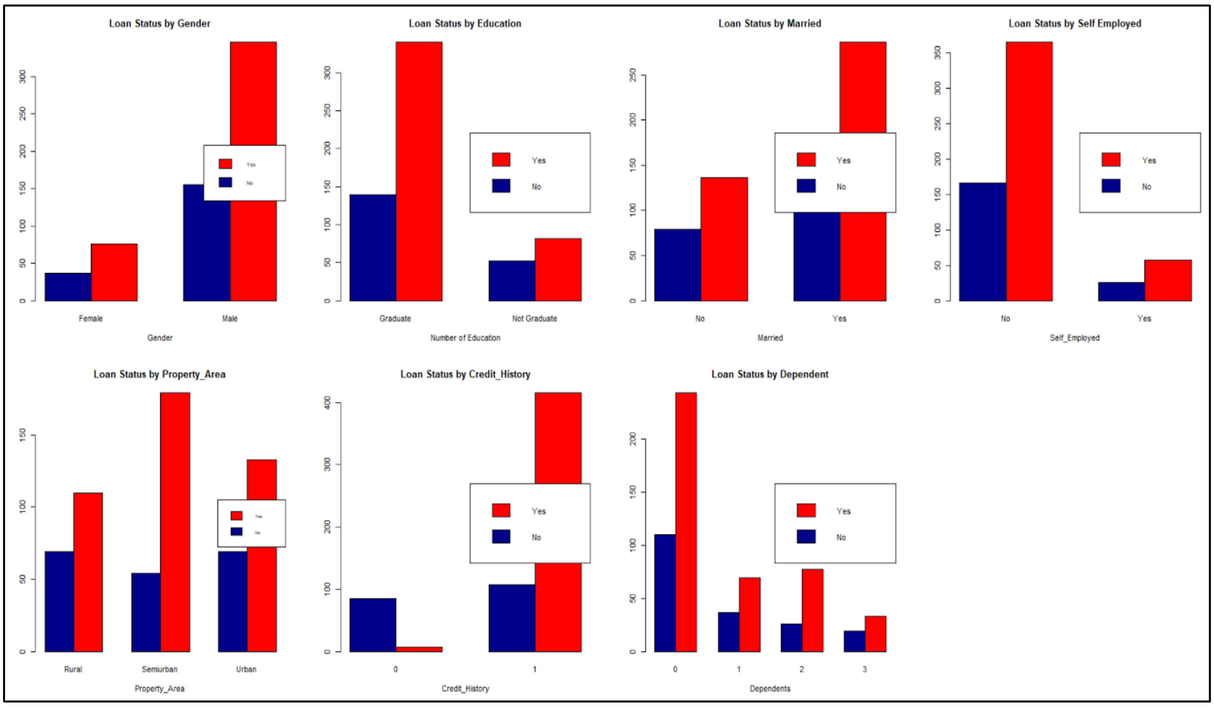


Figure : Each categorical variable was compared with loans status

The bar charts showed that each of the categorical variables were not showing an obvious relationship with loan status independently except the Credit_History variable. The number of rejected loan status was higher if the applicant did not have any credit history. This showed that the Credit History was the most significant variable that will affect the loan status. However, all the features were assumed to be important in determining loan status as each of the feature was correlated with each other. For example, if applicant is a male, educated, married, self-employed, has credit history will have higher chance to successfully to apply loan. Hence, significant features would be decided during each model. Before each model were trained, the clean data were split into train and test data with the ratio of 70 percent of clean data was train data and 30 percent of clean data was test data.

RESULTS AND DISCUSSION

Decision Tree Model

In this model, the prediction of the loan approval was successfully done by using the supervised learning model which is decision tree model. From the result of this model, some features were used to predict the customers can successfully get the loan or not. For example, Gender, Married, Dependents, Education, Self_Employed, Loan_Amount_Term, Credit_History, Property_Area, logtotalIncome, logLoanAmount are the features used to build this model. This model was trained by classification method with the features listed above.

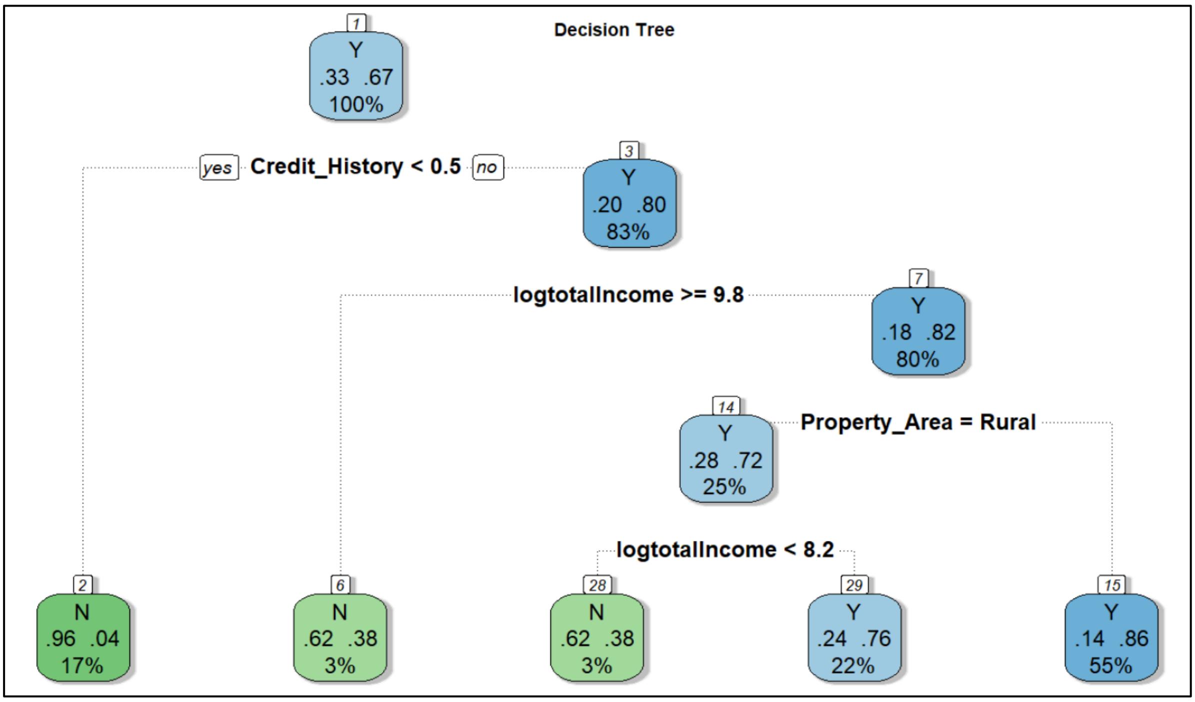
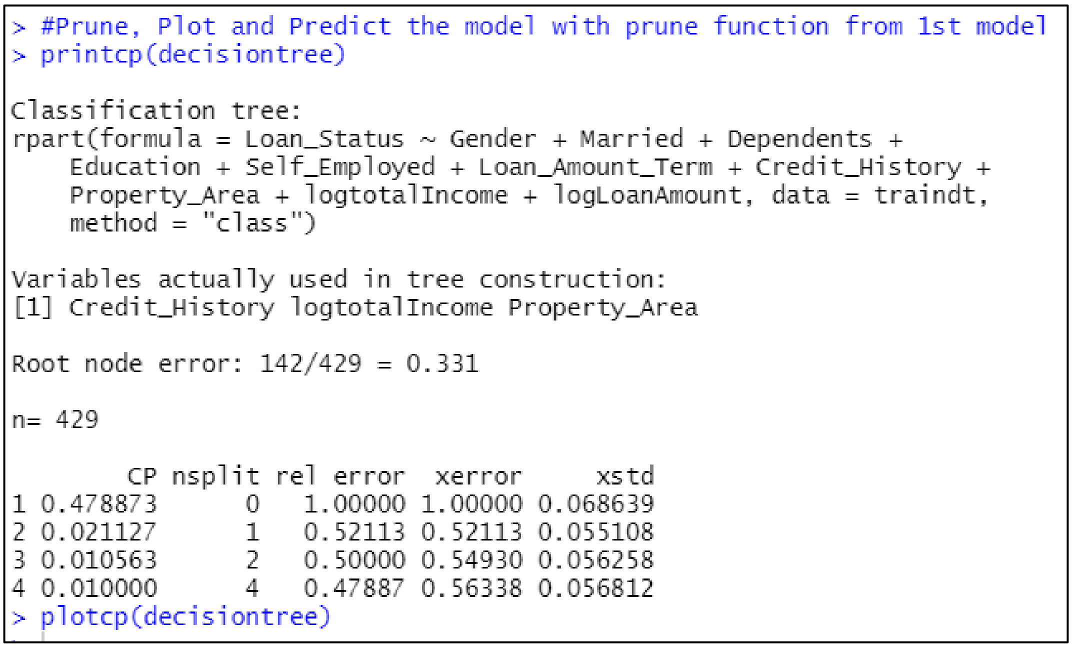


Figure 1.1: Plot of Decision Tree

Figure 1 was plotted by using *fancyRpartPlot()* function. This figure shows that the Loan_Status at the root node which is at the top of the tree is 67% of overall probability that applicants will get the loan and 33% percentage failed to get the loan in overall probability. Next, if the Credit_History is less than 0.5, there is only 4% of chances to get the loan and 17% of applicants in this dataset failed to get the loan with Credit_History less than 0.5. If the Credit_History is more than 0.5, it will go to another decision node which have the feature of logtotalIncome. From this figure, the highest probability to get the loan should have the features, which are the Credit_History more than 0.5, logtotalIncome less than or equal 9.8, Property_Area not at the rural area. With these features, the applicants will have the probability 86% successfully get the loan. In this data set, there is 55% of applicants fulfilled the condition and successfully get the loan.



```
> #Prune, Plot and Predict the model with prune function from 1st model
> printcp(decisiontree)

Classification tree:
rpart(formula = Loan_Status ~ Gender + Married + Dependents + 
    Education + Self_Employed + Loan_Amount_Term + Credit_History + 
    Property_Area + logtotalIncome + logLoanAmount, data = traindt, 
    method = "class")

Variables actually used in tree construction:
[1] Credit_History logtotalIncome Property_Area 

Root node error: 142/429 = 0.331

n= 429 

        CP nsplit rel error  xerror     xstd
1 0.478873      0   1.00000 1.00000 0.068639
2 0.021127      1   0.52113 0.52113 0.055108
3 0.010563      2   0.50000 0.54930 0.056258
4 0.010000      4   0.47887 0.56338 0.056812
> plotcp(decisiontree)
```

Figure 1.2: Content of CP table

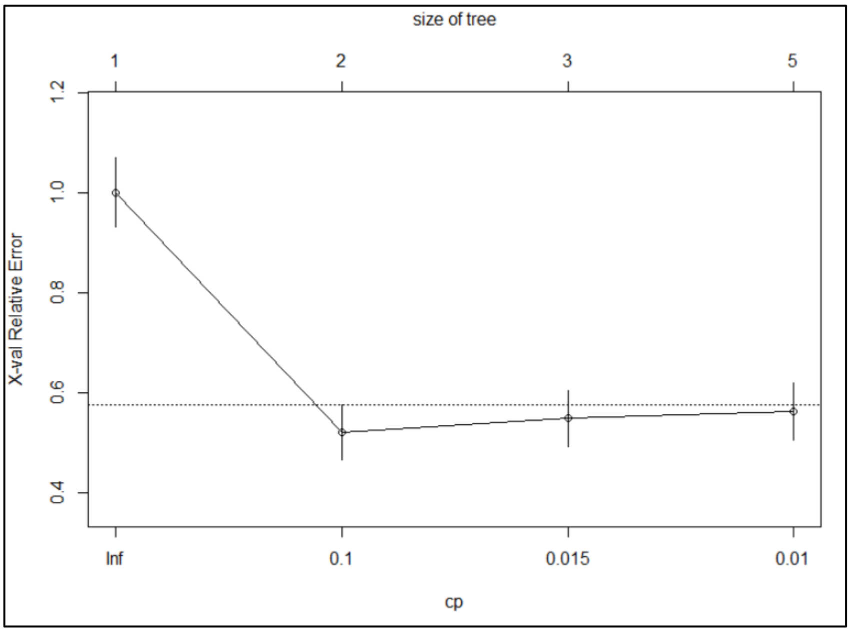


Figure 1.3: Plot of CP Graph

Due to the size of the tree are too large in the Figure 1.1 and some features were less effective to the model's accuracy, therefore prune the tree is required in this model. To choose the pruned tree size, smallest cross-validated error (xerror) needs to be chosen. From the figure 1.2, printcp() function is used to show the Complexity Parameter (cp) table. In the cp table, the lowest xerror is 0.52113 which had the cp, 0.021127 in size 2. To confirm the size of the tree, plotcp() function is used also to show the xerror graph. In the graph in figure 1.3, the lowest xerror had the size 2. Therefore, size of the tree is chosen as 2.

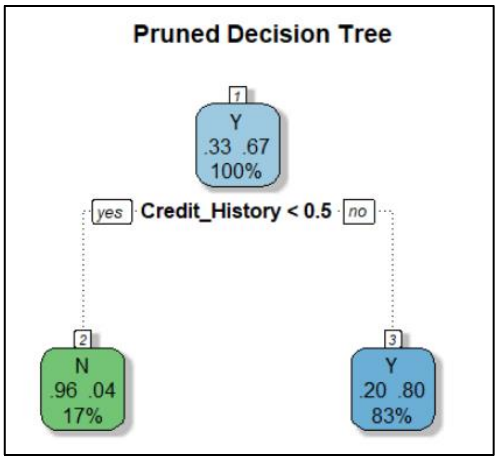
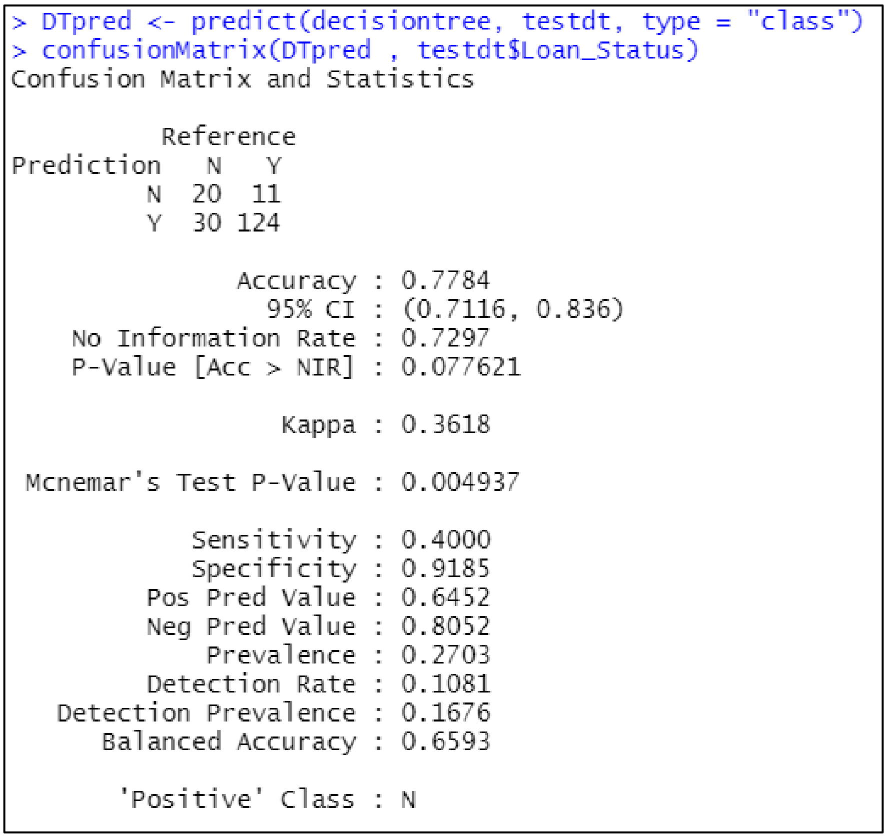


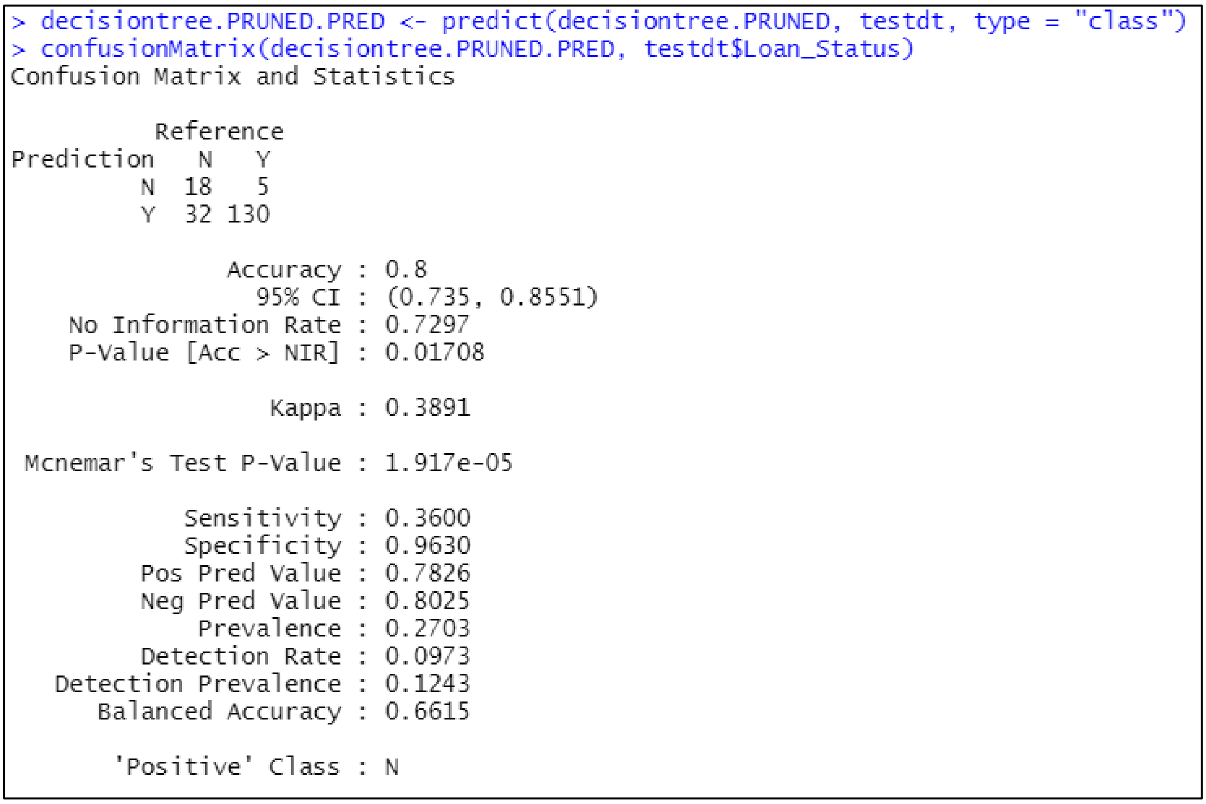
Figure 1.4: Plot of Pruned Decision Tree

After the tree is pruned, a new decision tree is plotted and it shows only one feature in the tree which is the most significant feature for this model. From this tree, Credit_History less than 0.5 will only have 4% probability to get the loan. For the Credit_History more than 0.5, will have 80% of chances to get the loan.



```
> DTpred <- predict(decisiontree, testdt, type = "class")
> confusionMatrix(DTpred , testdt$Loan_Status)
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N  20  11
         Y  30 124

               Accuracy : 0.7784
                 95% CI : (0.7116, 0.836)
    No Information Rate : 0.7297
    P-Value [Acc > NIR] : 0.077621

                  Kappa : 0.3618

 Mcnemar's Test P-Value : 0.004937

            Sensitivity : 0.4000
            Specificity : 0.9185
         Pos Pred Value : 0.6452
         Neg Pred Value : 0.8052
             Prevalence : 0.2703
         Detection Rate : 0.1081
   Detection Prevalence : 0.1676
      Balanced Accuracy : 0.6593

       'Positive' Class : N
```

Figure 1.5: Confusion Matrix of Prediction 1



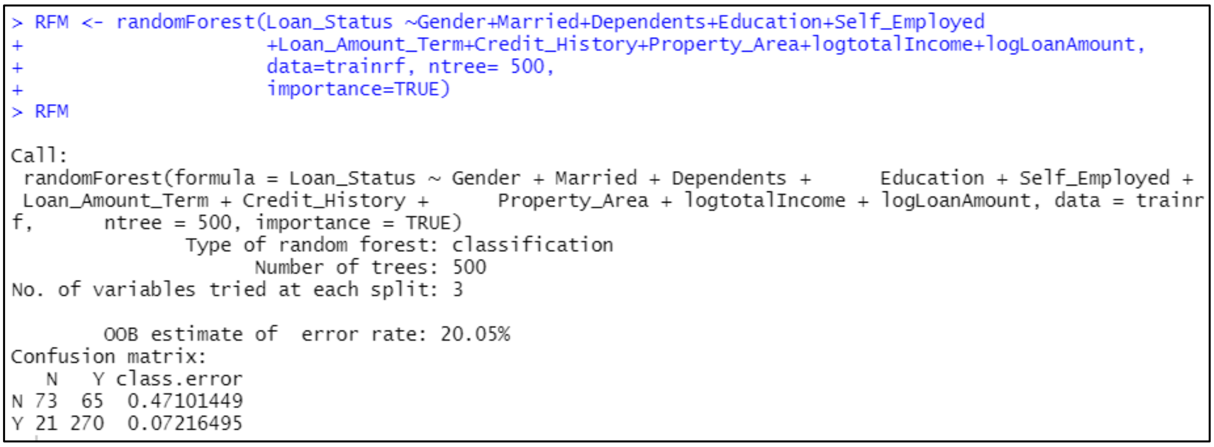
```
> decisiontree.PRUNED.PRED <- predict(decisiontree.PRUNED, testdt, type = "class")
> confusionMatrix(decisiontree.PRUNED.PRED, testdt$Loan_Status)
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N  18   5
         Y  32 130

               Accuracy : 0.8
                 95% CI : (0.735, 0.8551)
    No Information Rate : 0.7297
    P-Value [Acc > NIR] : 0.01708

                  Kappa : 0.3891

 Mcnemar's Test P-Value : 1.917e-05

            Sensitivity : 0.3600
            Specificity : 0.9630
         Pos Pred Value : 0.7826
         Neg Pred Value : 0.8025
             Prevalence : 0.2703
         Detection Rate : 0.0973
   Detection Prevalence : 0.1243
      Balanced Accuracy : 0.6615

       'Positive' Class : N
```

Figure 1.6: Confusion Matrix of Prediction 2

Prediction is done before the tree is pruned and after the tree is pruned. With the use of confusionMatrix() function, the accuracy of the prediction before the tree pruned is 79.46% in Figure 1.5 and the accuracy of the prediction after the tree pruned is 80% in Figure 1.6. This shows that this decision tree model is more accurate after the tree is pruned.

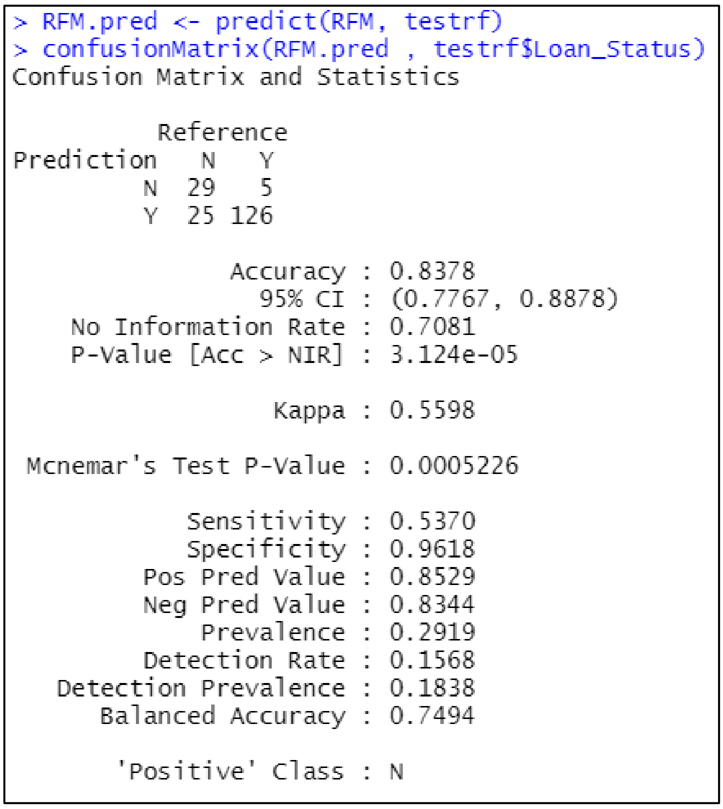
Random Forest Model

In the project, a predictive model was built by using the random forest classification to predict the Loan_Status variable using the other 10 variables including Gender, Married, Dependents, Education, Self_Employes, Loan_Amount_Term, Credit_History, Property_Area, logLoanAmount, and logtotalIncome in the training data. The model was fitted with 429 sample observations and was built by growing 500 trees and the number of variables utilised at each node was set to its default amount of sqrt(number of variables).



```
> RFM <- randomForest(Loan_Status ~Gender+Married+Dependents+Education+Self_Employed
+                     +Loan_Amount_Term+Credit_History+Property_Area+logtotalIncome+logLoanAmount,
+                     data=trainrf, ntree= 500,
+                     importance=TRUE)
> RFM

Call:
 randomForest(formula = Loan_Status ~ Gender + Married + Dependents +      Education + Self_Employed +
 Loan_Amount_Term + Credit_History +      Property_Area + logtotalIncome + logLoanAmount, data = trainr
f,      ntree = 500, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 3

        OOB estimate of  error rate: 20.05%
Confusion matrix:
   N   Y class.error
N 73  65  0.47101449
Y 21 270  0.07216495
```

Figure 2

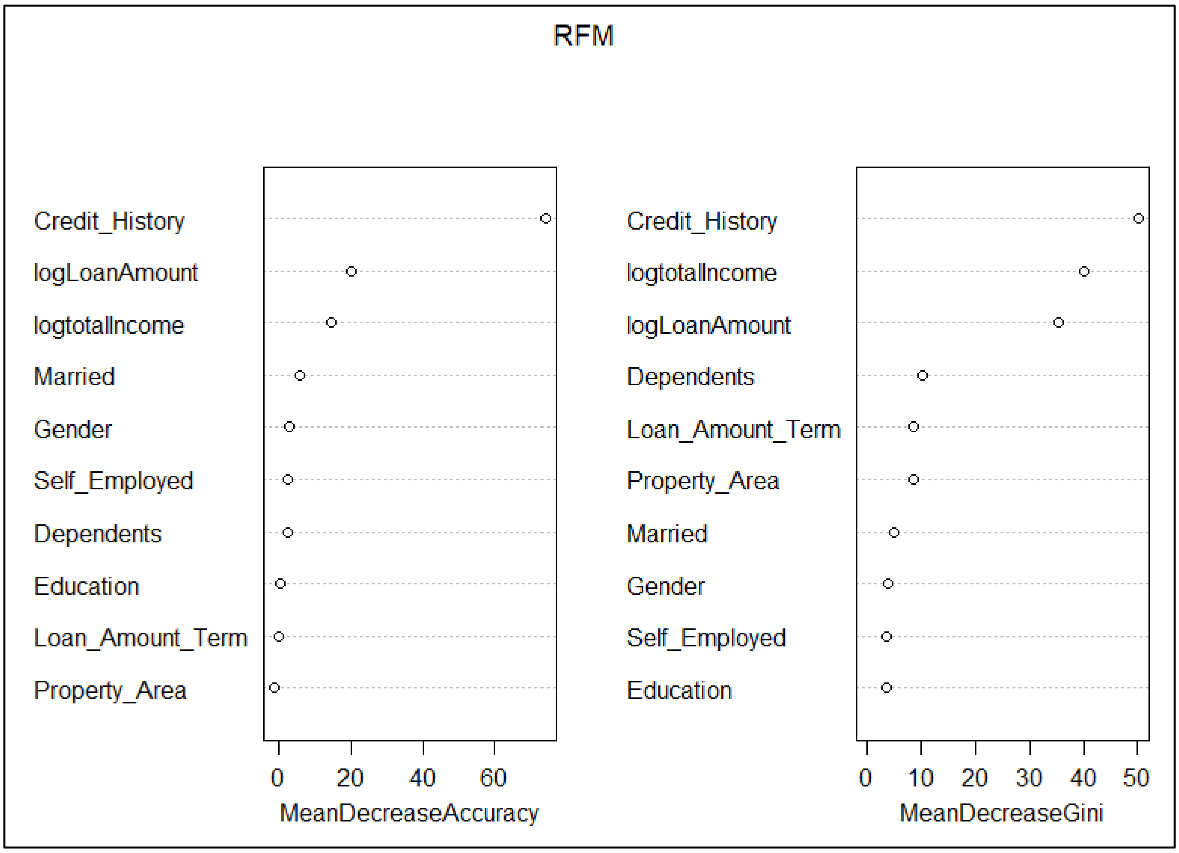


```
> RFM.pred <- predict(RFM, testrf)
> confusionMatrix(RFM.pred , testrf$Loan_Status)
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N  29   5
         Y  25 126

               Accuracy : 0.8378
                 95% CI : (0.7767, 0.8878)
    No Information Rate : 0.7081
    P-Value [Acc > NIR] : 3.124e-05

                  Kappa : 0.5598

 Mcnemar's Test P-Value : 0.0005226

            Sensitivity : 0.5370
            Specificity : 0.9618
         Pos Pred Value : 0.8529
         Neg Pred Value : 0.8344
             Prevalence : 0.2919
         Detection Rate : 0.1568
   Detection Prevalence : 0.1838
      Balanced Accuracy : 0.7494

       'Positive' Class : N
```

Figure

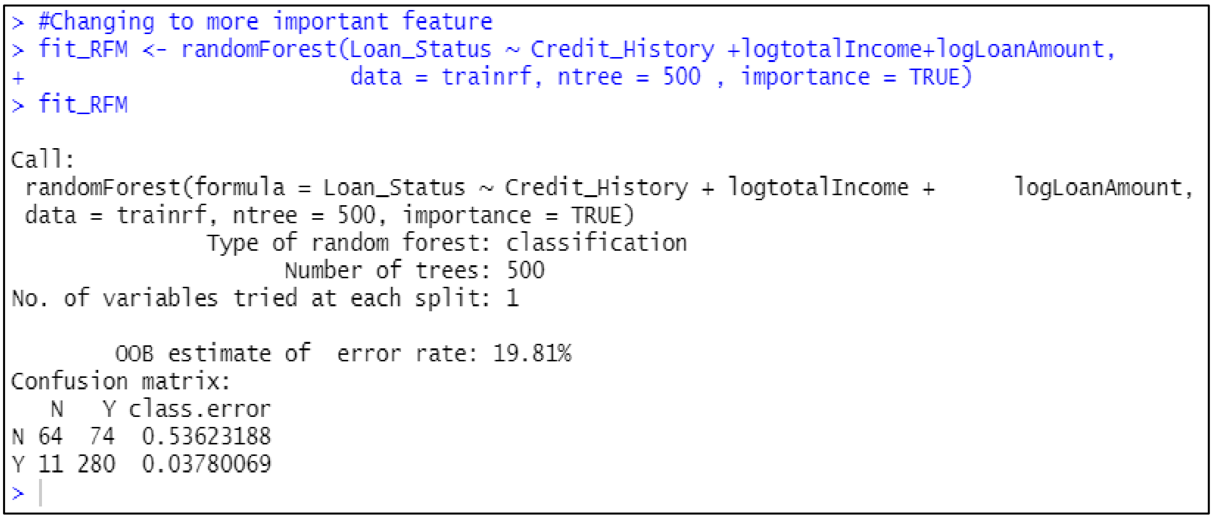
The model's contents are printed, and it reveals that the model produces an OOB estimate of error rate of 20.05 % , indicating that there is 20.05 percent of miss-classification error in the total number of observations in training data. In addition, the contents include a confusion table, which demonstrates that the model has 343 accurately classified data. The prediction was then applied to a test sample of 185 observations to validate the model's forecasting strength. A confusion matrix was given to describe the performance of the classification model by comparing predicted and actual values in the testing data. The Confusion Matrix shows that the accuracy of the model for the test sample is 83.78%.

Next, the dotcharts of MeanDecreaseAccuracy and MeanDecreaseGini were then plotted using the varImpPlot() method. It is critical to utilise the variables important plot to determine which variables have the most influence on the random forest model's predictions, or which variables the random forest model is most reliant on.



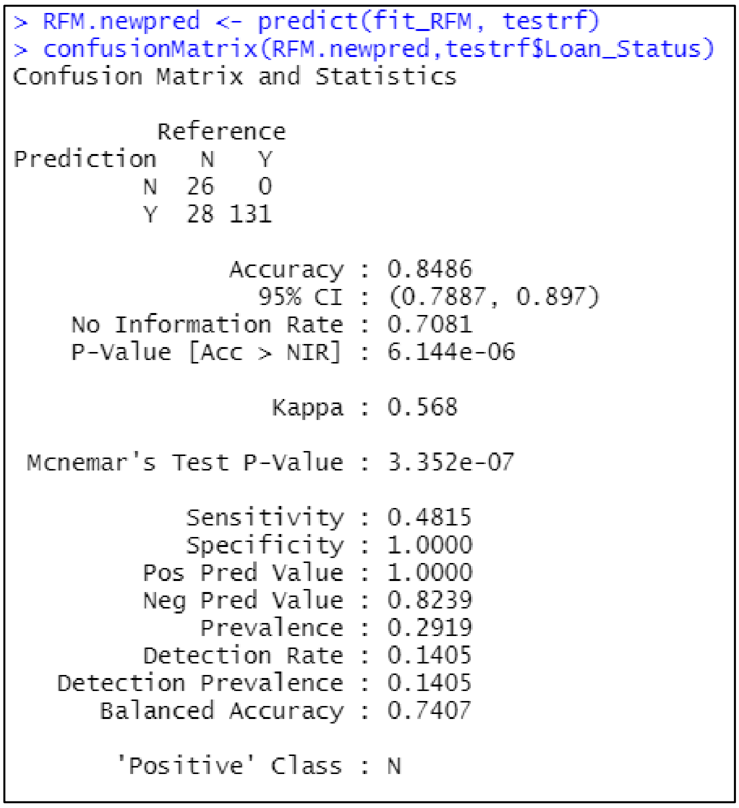
Figure

Based on the Figure, the variables Credit History, logLoanAmount, and logtotalIncome have the highest Mean Decrease Accuracy and Mean Decrease Gini in the Figure. This simply indicates that omitting any of the variables reduces both the prediction's accuracy and the purity of the nodes. As a result, the most relevant variables in predicting loan approval are Credit History, logLoanAmount, and logtotalIncome. Thus, these factors were used to create a new predictive model to study the performance change when the more correlated variables were chosen.



```
> #Changing to more important feature
> fit_RFM <- randomForest(Loan_Status ~ Credit_History +logtotalIncome+logLoanAmount,
+                         data = trainrf, ntree = 500 , importance = TRUE)
> fit_RFM

Call:
 randomForest(formula = Loan_Status ~ Credit_History + logtotalIncome +      logLoanAmount,
 data = trainrf, ntree = 500, importance = TRUE)
               Type of random forest: classification
                     Number of trees: 500
No. of variables tried at each split: 1

        OOB estimate of  error rate: 19.81%
Confusion matrix:
   N   Y class.error
N 64  74  0.53623188
Y 11 280  0.03780069
> 
```

Figure



```
> RFM.newpred <- predict(fit_RFM, testrf)
> confusionMatrix(RFM.newpred,testrf$Loan_Status)
Confusion Matrix and Statistics

          Reference
Prediction   N   Y
         N  26   0
         Y  28 131
                                          
               Accuracy : 0.8486          
                 95% CI : (0.7887, 0.897) 
    No Information Rate : 0.7081          
    P-Value [Acc > NIR] : 6.144e-06       
                                          
                  Kappa : 0.568           
                                          
 Mcnemar's Test P-Value : 3.352e-07       
                                          
            Sensitivity : 0.4815          
            Specificity : 1.0000          
         Pos Pred Value : 1.0000          
         Neg Pred Value : 0.8239          
             Prevalence : 0.2919          
         Detection Rate : 0.1405          
   Detection Prevalence : 0.1405          
      Balanced Accuracy : 0.7407          
                                          
       'Positive' Class : N               
```

Figure

By choosing the more important variables, the OOB estimate of the new model's error rate for the training sample dropped to 19.81% and the prediction accuracy for the testing sample was increased to 84.86% higher than the 83.78% by the previous model. According to the result above, it is clear that the variables giving the largest contributions to the loan approval predicting model are Credit History, logLoanAmount, and logtotalIncome. This demonstrates the need of selecting more correlated variables in order to improve model's performance.

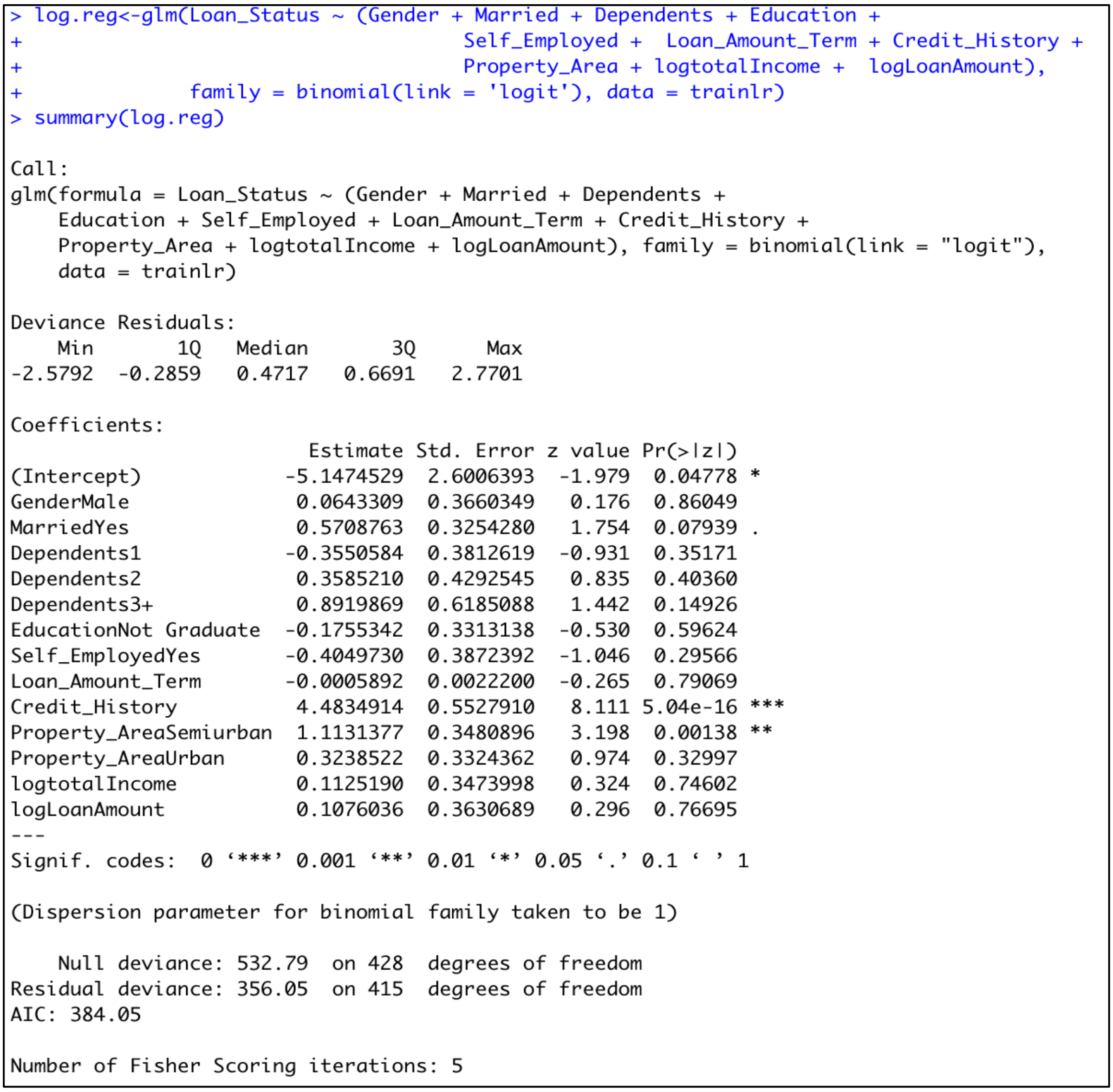
Logistic Regression Model

Logistic regression is one of the statistical techniques in machine learning used to form predictive models. It is one of the most popular classification algorithms mostly used for binary classification problems. Therefore, it is essential to have a good grasp on logistic regression algorithms. Logistic regression is used when the dependent variable is categorical. For example is to predict whether a loan will be approved (1) or rejected (0).

After all the pre-processing of data had been done, a logistic regression model could start to be created. Followed by this, the model was trained using the fit method that contains 70% of the dataset. This will be a binary classification model. Firstly, the importance of the variables was examined logically to choose a better variable for analysis. Thus, a result of the chances that an applicant's application would be approved is higher if:

- Applicants took a loan before. Credit history is the variable which answers that.
- Applicants with higher incomes
- Applicants with higher education
- Applicants who have stable job

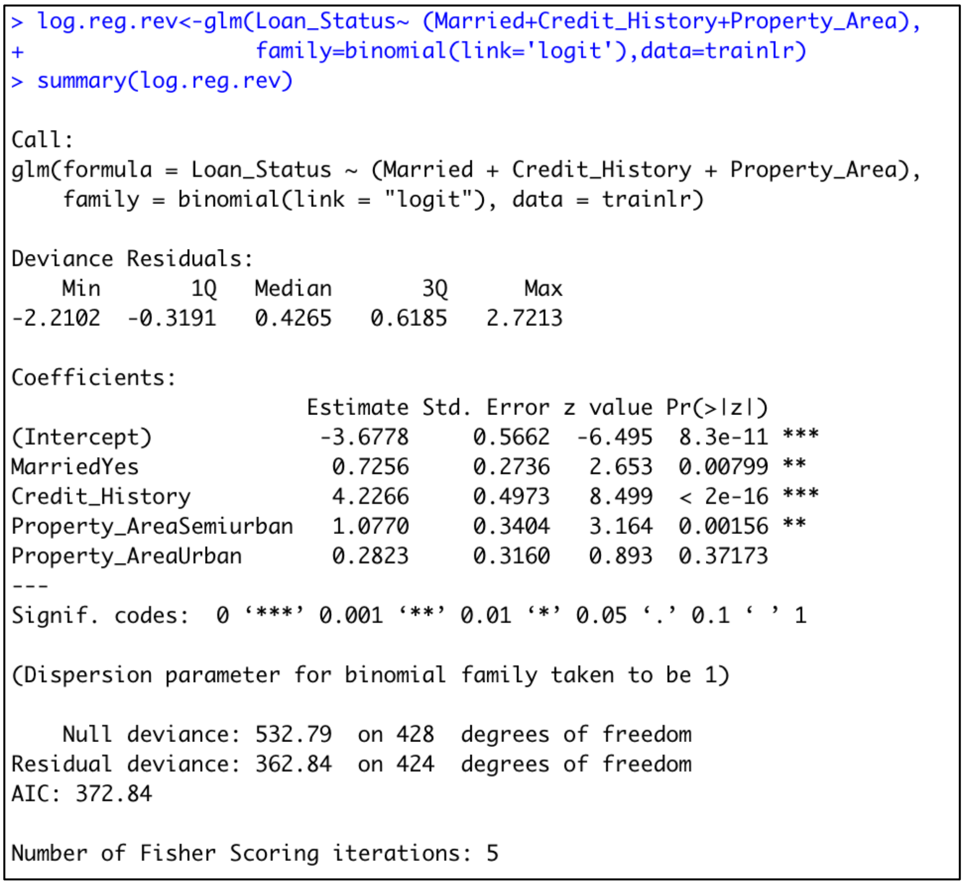
Two Generalized Linear Model (GLM) were built to determine which would be more optimal in predicting loan approval. The first model was built using all predictors in the dataset against the dependent variables as shown below. When the first model was built, predictor variables such as gender, loan amout term, loan amount and self employed were not statistically significant. Keeping them in the model may contribute to overfitting. The AIC score was 384.05 on 415 degrees of freedom. Pr(>|z|) could be explained as the "p-value" of the test for whether the coefficient point estimate was significantly different from 0. Intuitively, it told if the point estimate had been calculated precisely enough to distinguish it from zero and "precisely enough" was defined using p-value (<0.05). As shown below, there were 3 predictors generated from this model that were considered as important variables by looking at the "p-value". Thus, these 3 predictors were used to build the second model to improve the performance of the model and to reduce the computational cost of modelling.



```
> log.reg<-glm(Loan_Status ~ (Gender + Married + Dependents + Education +
+                                      Self_Employed +  Loan_Amount_Term + Credit_History +
+                                      Property_Area + logtotalIncome +  logLoanAmount),
+               family = binomial(link = 'logit'), data = trainlr)
> summary(log.reg)

Call:
glm(formula = Loan_Status ~ (Gender + Married + Dependents +
    Education + Self_Employed + Loan_Amount_Term + Credit_History +
    Property_Area + logtotalIncome + logLoanAmount), family = binomial(link = "logit"),
    data = trainlr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.5792  -0.2859   0.4717   0.6691   2.7701

Coefficients:
                         Estimate Std. Error z value Pr(>|z|)
(Intercept)            -5.1474529  2.6006393  -1.979  0.04778 *
GenderMale              0.0643309  0.3660349   0.176  0.86049
MarriedYes              0.5708763  0.3254280   1.754  0.07939 .
Dependents1            -0.3550584  0.3812619  -0.931  0.35171
Dependents2             0.3585210  0.4292545   0.835  0.40360
Dependents3+            0.8919869  0.6185088   1.442  0.14926
EducationNot Graduate  -0.1755342  0.3313138  -0.530  0.59624
Self_EmployedYes       -0.4049730  0.3872392  -1.046  0.29566
Loan_Amount_Term       -0.0005892  0.0022200  -0.265  0.79069
Credit_History          4.4834914  0.5527910   8.111 5.04e-16 ***
Property_AreaSemiurban  1.1131377  0.3480896   3.198  0.00138 **
Property_AreaUrban      0.3238522  0.3324362   0.974  0.32997
logtotalIncome          0.1125190  0.3473998   0.324  0.74602
logLoanAmount           0.1076036  0.3630689   0.296  0.76695
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 532.79  on 428  degrees of freedom
Residual deviance: 356.05  on 415  degrees of freedom
AIC: 384.05

Number of Fisher Scoring iterations: 5
```

Figure: Summary of glm() function for all predictors variables

The second GLM was built by using only 3 predictors which were considered as important variables. The AIC score was 372.84 which did decrease, therefore indicating that the performance of the model had improved. Thus, the performance of the second model was better than the first model, the first model had been rejected and the second model will be kept as the optimal model going forward.



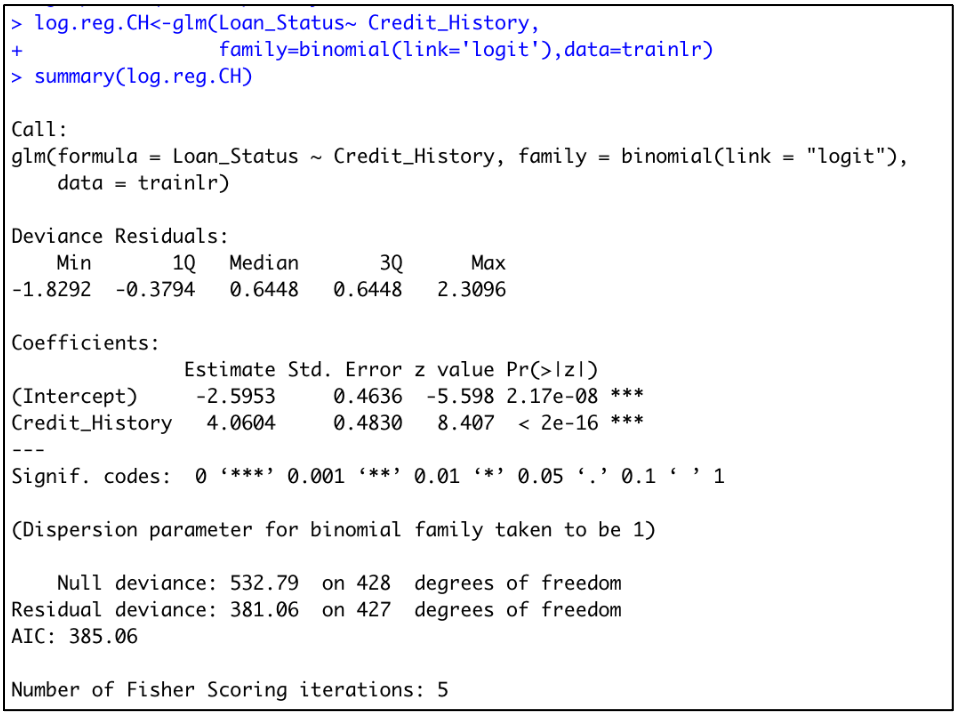
```
> log.reg.rev<-glm(Loan_Status~ (Married+Credit_History+Property_Area),
+                  family=binomial(link='logit'),data=trainlr)
> summary(log.reg.rev)

Call:
glm(formula = Loan_Status ~ (Married + Credit_History + Property_Area),
    family = binomial(link = "logit"), data = trainlr)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-2.2102  -0.3191   0.4265   0.6185   2.7213

Coefficients:
                       Estimate Std. Error z value Pr(>|z|)
(Intercept)             -3.6778     0.5662  -6.495  8.3e-11 ***
MarriedYes               0.7256     0.2736   2.653  0.00799 **
Credit_History           4.2266     0.4973   8.499  < 2e-16 ***
Property_AreaSemiurban   1.0770     0.3404   3.164  0.00156 **
Property_AreaUrban       0.2823     0.3160   0.893  0.37173
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 532.79  on 428  degrees of freedom
Residual deviance: 362.84  on 424  degrees of freedom
AIC: 372.84

Number of Fisher Scoring iterations: 5
```

Figure: summary of glm() function for significant variables

After using the stepwise regression function, there were 3 predictors remaining that were considered strong enough to predict whether or not a person could be approved for a loan. From the looks of the model, Credit_History with a positive coefficient of 4.2266 had the most impact on the target variable while having the smallest p-value. This meant that an increase in a good credit history was associated with an increased chance of getting approved for a loan. This made sense as many of the finance companies tend to look at the credit history to determine how responsible people were.

Hence, the Credit_History variable had been chosen as the most important predictor in the logistic regression model in this project. As could be seen below, a low p-value was obtained from Credit_History, thus it was the most significant variable among all the variables.

As mentioned before, the Credit_History variable had been chosen in the logistic regression model in this project. It played a major role in determining who gets approved for a loan. It would be the first thing the lender looks at when assessing the potential borrower’s qualification. A model with Credit_History predictor variable was built to figure out the performance of the model with the most significant predictor. With Credit_History being the most significant predictor, the probability of being approved based on this predictor had also been shown below.



```
> log.reg.CH<-glm(Loan_Status~ Credit_History,
+                 family=binomial(link='logit'),data=trainlr)
> summary(log.reg.CH)

Call:
glm(formula = Loan_Status ~ Credit_History, family = binomial(link = "logit"), 
    data = trainlr)

Deviance Residuals: 
    Min       1Q   Median       3Q      Max  
-1.8292  -0.3794   0.6448   0.6448   2.3096  

Coefficients:
               Estimate Std. Error z value Pr(>|z|)    
(Intercept)     -2.5953     0.4636  -5.598 2.17e-08 ***
Credit_History   4.0604     0.4830   8.407  < 2e-16 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 532.79  on 428  degrees of freedom
Residual deviance: 381.06  on 427  degrees of freedom
AIC: 385.06

Number of Fisher Scoring iterations: 5
```

Figure: summary of glm() function for Credit History feature

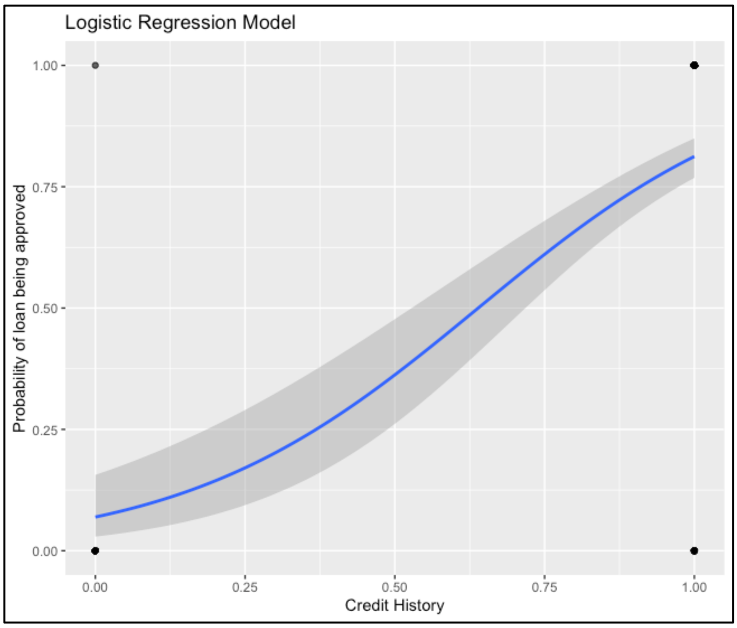


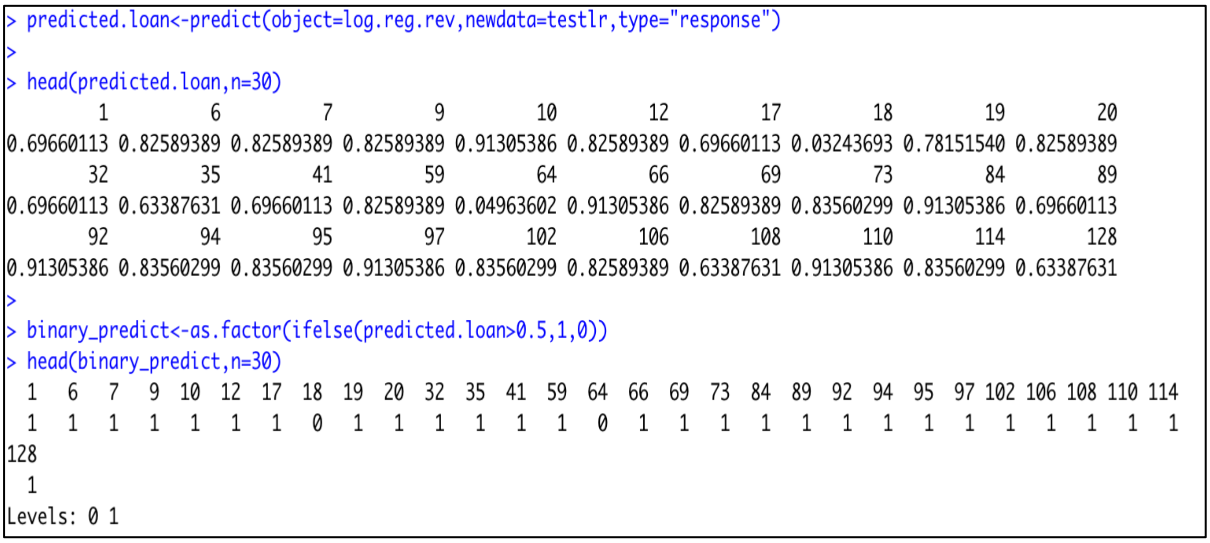
Figure: Graph for the relationship between the probability of loan being approved and credit history

When fitting the line to the points, a slight a S-shape curve is produced.

Among these models, the model with 3 significant predictors variables was chosen in this project to validate with the test dataset. This is due the AIC in the model with 3 significant predictors variables was the lowest relatively. The lowest AIC indicates that the model contributed the best performance in any unseen data.

Prediction

Next, the model was tested with our test dataset which is 30% splitted from the loan dataset to make predictions. predict() function was used to generate predictions from the test dataset for our logistic regression model.



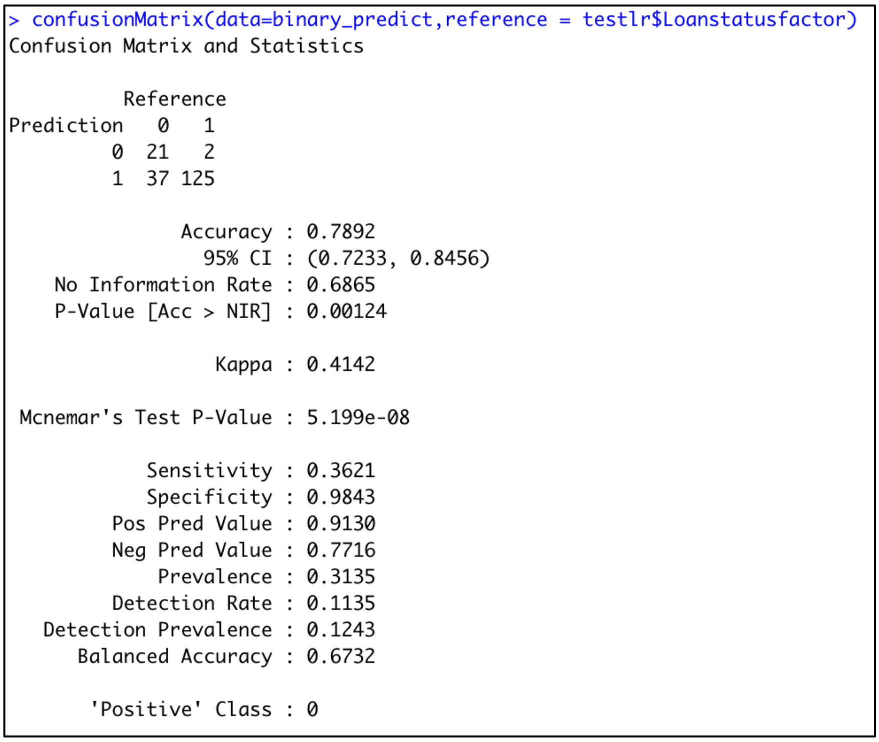
```
> predicted.loan<-predict(object=log.reg.rev,newdata=testlr,type="response")
>
> head(predicted.loan,n=30)
         1          6          7          9         10         12         17         18         19         20 
0.69660113 0.82589389 0.82589389 0.82589389 0.91305386 0.82589389 0.69660113 0.03243693 0.78151540 0.82589389 
        32         35         41         59         64         66         69         73         84         89 
0.69660113 0.63387631 0.69660113 0.82589389 0.04963602 0.91305386 0.82589389 0.83560299 0.91305386 0.69660113 
        92         94         95         97        102        106        108        110        114        128 
0.91305386 0.83560299 0.83560299 0.91305386 0.83560299 0.82589389 0.63387631 0.91305386 0.83560299 0.63387631 
>
> binary_predict<-as.factor(ifelse(predicted.loan>0.5,1,0))
> head(binary_predict,n=30)
  1   6   7   9  10  12  17  18  19  20  32  35  41  59  64  66  69  73  84  89  92  94  95  97 102 106 108 110 114 
  1   1   1   1   1   1   1   0   1   1   1   1   1   1   0   1   1   1   1   1   1   1   1   1   1   1   1   1   1 
128 
  1 
Levels: 0 1
```

Figure: The probability for randomised 30 applicants' loans being approved

The probability for every applicant in test dataset was calculated by using the predict() function. 0.50 was set as the threshold value. Hence, if the probability that greater than threshold will be classified as approved loan while if less than the threshold will be classified as rejected loan.

Model Evaluation

After building a logistic regression model and making predictions, the accuracy for the model to make predictions for our test dataset was calculated by using a confusion matrix. Confusion matrix was used to check whether the model generates right or false predictions with respect to the actual value from the test dataset.

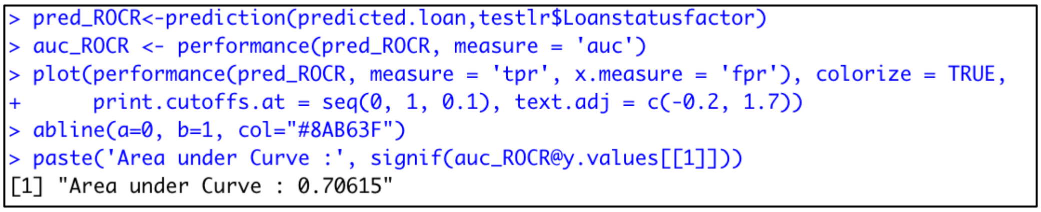


```
> confusionMatrix(data=binary_predict,reference = testlr$Loanstatusfactor)
Confusion Matrix and Statistics

          Reference
Prediction   0   1
         0  21   2
         1  37 125
                                          
               Accuracy : 0.7892          
                 95% CI : (0.7233, 0.8456)
    No Information Rate : 0.6865          
    P-Value [Acc > NIR] : 0.00124         
                                          
                  Kappa : 0.4142          
                                          
 Mcnemar's Test P-Value : 5.199e-08       
                                          
            Sensitivity : 0.3621          
            Specificity : 0.9843          
         Pos Pred Value : 0.9130          
         Neg Pred Value : 0.7716          
             Prevalence : 0.3135          
         Detection Rate : 0.1135          
   Detection Prevalence : 0.1243          
      Balanced Accuracy : 0.6732          
                                          
       'Positive' Class : 0               
```

Figure: Confusion matrix for logistic regression model with 3 significant variables

Regarding the result of the confusion matrix, the accuracy of the model measured is 0.7892. The model correctly predicts from the test dataset that 125 loan ID (true positive) is able to pay their loan while 21 loan ID default (true negative). The overall accuracy was calculated by $\frac{\text{correctprediction}}{\text{totalobservations}}$ or equal to $\frac{(21+125)}{(21+2+37+125)}$ which gave a result of 0.7892. That means that the model (logistic regression) had correctly predicted 146 outcomes out of 185 observations. However, the confusion matrix suggested that the model had False Positive of 2 data which indicates that the model predicted the loan for 2 applicants will rejected but they actually got approved loan. However, there were also 37 False Negative data which implies that the model predicted that the loan for 37 applicants will approved but actually their loan got rejected.

Other than confusion matrix, ROC graph was also plotted to examine the performance of the classifier used which is logistic regression. The accuracy of the classifier was calculated by measuring the area under the ROC graph.



```
> pred_ROCR<-prediction(predicted.loan,testlr$Loanstatusfactor)
> auc_ROCR <- performance(pred_ROCR, measure = 'auc')
> plot(performance(pred_ROCR, measure = 'tpr', x.measure = 'fpr'), colorize = TRUE,
+      print.cutoffs.at = seq(0, 1, 0.1), text.adj = c(-0.2, 1.7))
> abline(a=0, b=1, col="#8AB63F")
> paste('Area under Curve :', signif(auc_ROCR@y.values[[1]]))
[1] "Area under Curve : 0.70615"
```

Figure: coding and area for the ROC graph

AUC calculated is 0.70615.

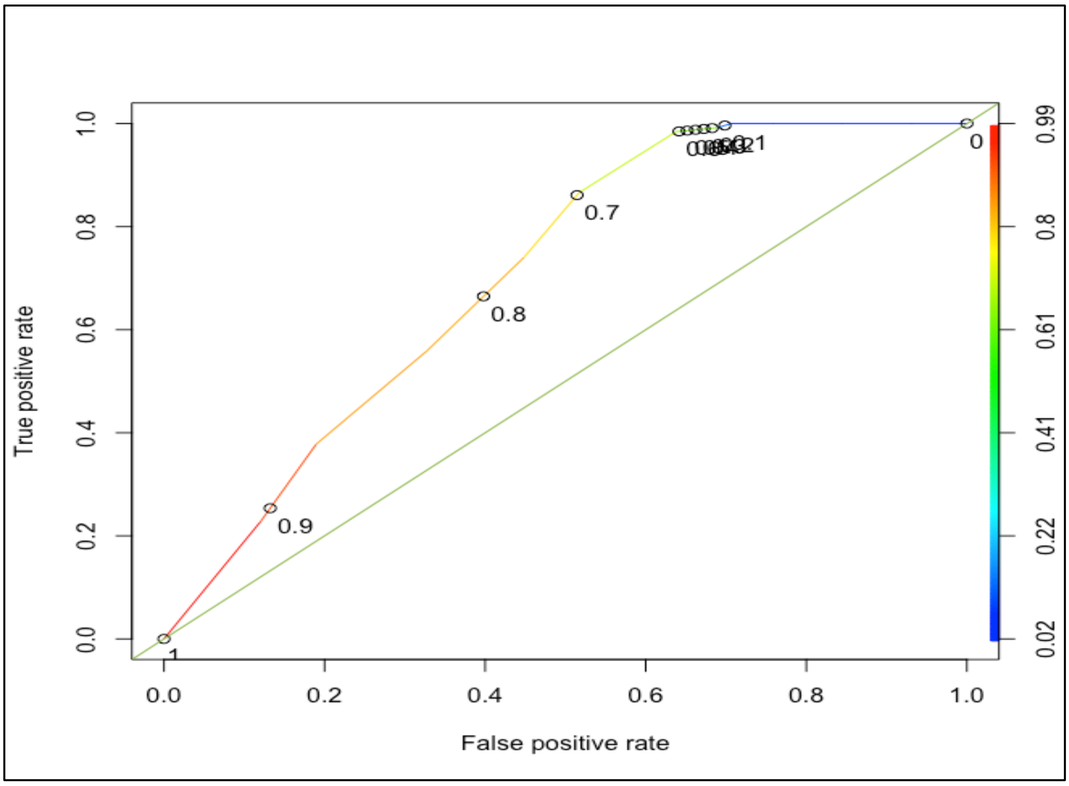


Figure: ROC curve for logistic regression model

The ROC curve plots the False Positive Rate (FPR) on the X-axis and the True Positive Rate (TPR) on the Y-axis for all possible thresholds (or cut off values).

True Positive Rate (TPR) or sensitivity: the proportion of actual positives that are correctly identified as such.

True Negative Rate (TNR) or specificity: the proportion of actual negatives that are correctly identified as such.

False Positive Rate (FPR) or 1-specificity: the proportion of actual negatives that are wrongly identified as positives.

The ROC curves were plotted to display the trade-off between True Positive Rate (sensitivity) and False Positive Rate (specificity).

The green line was represented a completely uninformative test which corresponds to an AUC of 0.5. A curve pulled close to the upper left corner indicates a better performing test. On the other hand, an AUC that is close to 1 can also indicates a better performing test. So, the logistic regression model was performing well in this test.

Discussion

In decision tree model, it showed that the most important feature in this model is Credit_History. The first plotting of decision tree showed that there are few branches in the tree. The second plotting of the decision tree is the tree that had been pruned from the first tree, it showed that there is only one branch, and the decision node is Credit_History. By comparision between two plotting, the Credit_History is the only one that left after pruning the tree. Therefore, the Credit_History is the most important feature in the decision tree model.

In random forest model, varImpPlot() was used to determine the important features that would highly affect the loan status. VarImpPlot() showed that credit history variable, loan amount and total income were the most significant features in the dataset. This is because these variables were the top 3 variables in the mean decrease accuracy plot and mean decrease Gini plot. Therefore, removing 3 of these variables will decrease the model's accuracy drastically and also reduce the model performance.

In logistic regression model, glm() was used to determine the important features that would highly affect the loan status. By looking at the p-value and AIC score, it showed that Maried, Credit_History and Property_Area were considered as important variables. While the AIC score decrease, indicating that the performance of the model had improved.

By comparision the accuracy between three models, the logistic regression model has the lowest accuracy which is 78.92%. The decision tree model has taken the middle position with the accuracy, 80.00%. The highest accuracy goes to the random forest model with 84.86%. Without comparing between the models, the three models that have been created are performing well.

From the result given by three models, the most important feature that will affect the accuracy in this project is Credit_History. The most suitable model to select the loan applicants is random forest model which gives 84.86% of accuracy.

OPPORTUNITIES

More data can provide the machine learning algorithms with more knowledge to understand the various situations. As a result, it can make comparisons before providing an answer. More training data can help to prevent biased decisions by aiming to include a wide range of data that covers a wide range of scenarios. With more data input, the model's accuracy can be improved. As more data is added, the likelihood of overfitting decreases rather than increases. It reduces the generalisation error because your model becomes more general as it is trained on more examples. Increasing the number of input features or columns may increase overfitting. This is because more features may be irrelevant or redundant, thus, there is more opportunity to complicate the model to fit the examples at hand. Double-descent occurs when the size of the training set is close to the number of model parameters.

The model can be improved by adjusting the feature selection and identifying the appropriate variables. Features have the greatest influence on the outcome and are one of the most important factors. Features have a large impact on a model's output and are one of the most important aspects of the model-building process. Finding the right variables or features to extract as new knowledge is the most effective way to improve ML model accuracy. You can correctly identify the most appropriate variables when you have a better understanding and visualisations. Before deploying a machine learning algorithm, it is critical to consider as many relevant variables and potential outcomes as possible.

By combining individual models, the ensemble method can improve prediction output. It is a popular method that is frequently used by combining multiple models to improve precision with bagging and boosting, such as the Random Forest technique. Ensemble methods are techniques for developing multiple models and then combining them to produce better results. Ensemble methods typically yield more accurate results than a single model. When compared to other more traditional methods, this method is quite complex, but it can produce highly accurate results. An ensemble can outperform any single contributing model in terms of prediction and performance.

Furthermore, the spread or dispersion of predictions as well as model performance can be reduced by ensemble technique. This technique is used to achieve better predictive performance than a single predictive model on a predictive modelling problem. This is accomplished by the model reducing the variance component of the prediction error by adding bias. The robustness or reliability in a model's average performance is also improved by applying ensemble method and this is one of the significant and underappreciated advantage of ensemble methods. Fitting the model multiple times on the training datasets and combining the predictions using a summary statistic, such as the mean for regression or the mode for classification, is the simplest ensemble. Importantly, because of the stochastic learning algorithm, differences in the composition of the training dataset, or differences in the model itself, each model must be slightly different. This will reduce the spread in the model's predictions. The mean performance will most likely be similar, but the worst- and best-case performance will be brought closer to the mean performance.

TRENDS

Machine learning is a branch in data science that studies the design of algorithms that can learn. The ultimate goal is to increase learning to the point where it becomes automatic, minimizing the need for humans to participate. Machine learning is now used in some form or another in almost every other tool and software on the internet. Machine Learning has become so prevalent that it is now the go-to method for businesses to handle a wide range of issues. The trend of machine learning will not be just what we see now and it has a lot of potential to contribute in many fields. Machine learning algorithms can process vast volumes of data and extract meaningful information using a variety of programming techniques. In this project, three models had been built which are decision tree, logistic regression and random forest based on loan approval prediction.

The first model that was used in this project is the decision tree. Decision tree is a flowchart-like structure in which each internal node represents a “test” on an attribute, each branch represents the outcome of the test, and each leaf node represents a class label. They are a versatile algorithm used to perform the tasks of classification and regression. According to research on loan approval prediction based on decision tree and random forest, with a confidence factor of 0.25 the accuracy is 63.39% using decision tree. Meanwhile, in the experiments without attribute selection using random forest the best accuracy is 85.75% (Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr. Suresh Kumar, 2020). They are using 12 attributes to do prediction and the results show that the accuracy of random forest is higher than decision tree.

The second model that was used in this project is logistic regression. It is applicable for categorical dependent variables using a given set of independent variables. Thus, the outcome must be a categorical or discrete value. The output can be either “yes” or “no”, “0” or “1”, “true” or “false”. In this project, the “S” shaped curve from the logistic function demonstrates the probability of loan approval. According to a research on loan approval prediction based on logistic regression, the model showed a precision of 0.8440 and 0.8244 with the train and test data respectively by considering 11 variables for the analysis of data. The performance of the model with both the train and test data was illustrated using a plot of train errors and test errors against sample size on the same axes. Thus, the study recommended the use of logistic regression in conjunction with supervised machine learning approach in loan default prediction in financial institutions (Mong'are, Dominic & Njoroge, Gladys & Muraya, Moses, 2019).

The third model that was used in this project is random forest. It is an ensemble learning method for classification, regression and other tasks that operates by constructing a multitude of decision trees at training time and outputting the class that is the mode of the classes or means prediction of the individual trees. According to a research on predicting loan default based on the random forest algorithm, the performance of random forest and decision tree have shown comparable performance than that of vector machine and logistic regression, but the random forest performs the best, with an accuracy of 98%, higher than the decision tree with an accuracy of 95%. They retained 15 attributes to do analysis. The precision and recall of the prediction model based on random forest are all above 0.95, indicating that the model has strong ability of generalization (Zhu, Lin & Qiu, Dafeng & Ergu, Daji & Ying, Cai & Liu, Kuiyi, 2019).

According to the research, it has been found that the overall best performance among decision tree, logistic regression and random forest was obtained by the supervised learning technique, random forest. Hence, the conclusion is the same as in this project, random forest has the highest sensitivity and precision while predicting loan approval. Thus, the random forest model appears to be a better option for such data. To predict loan approval, machine learning will be a trend in determining it. Taking out loans from financial institutions has become increasingly prevalent in today's environment. A big number of people apply for loans every day for a variety of reasons. However, all of these candidates are untrustworthy, and none of them can be approved. Making a loan approval decision carries a significant amount of risk. Thus, loan data should be collected from a variety of sources and apply machine learning techniques to extract useful information from it. Organizations can use this model to make the best judgment on whether to approve or deny a customer's loan request. This will be a trend in the future world to determine the credit worthiness of customers in order to protect the rights of each party.

CHALLENGES

In this project, some problems and challenges are needed to solve them. In the very beginning of doing the project, the topic of “Predicting the malware file” had been chosen but failed to do it because the data set is incomplete and not suitable to do analysis. Thus, the topic of this project had been changed to “Predict loan approval” and found a suitable data set to do analysis. Categorical variables and numerical variables need to be analysed when approaching the data set. Besides, data types, outliers, missing values and distributions of numerical and categorical variables are also needed to be identified. After these two steps had been done, it was found that the topic of this project needed to be changed. Due to time constraints, research on the new topic which is “Predict Loan Approval” needed to be done in a short time. Research and practise in this field should be done before the models and reports are started to do.

A package is a suitable way to organize the work and typically a package will include code, documentation for the package and the functions inside, some tests to check everything works as it should, and data sets. One of the challenges in this project was difficulty in understanding packages in R studio. It is because there are more than 10,000 packages available to be used. Each package is created to add specific functionality. Examples of packages that used in this project are plyr, VIM, mice and ggplot2. Before their specific function had been determined, the models failed to apply. For example, Plyr is an R package that makes it simple to split data apart, do stuff to it, and mash it back together. Thus, it is one of the challenges to understand which package is needed to apply in the models to carry out specific functions.

Furthermore, difficulties in understanding the code of model is also one of the challenges in this project. As everyone knows logistic regression is one of the most common machine learning algorithms used for binary classification. It predicts the probability of occurrence of a binary outcome using the logit function. A problem had been faced of plotting a sigmoid curve in logistic regression and what had been plotted was a linear line which didn’t meet the requirements. It showed the failure of understanding the model and the variables. Activation function(sigmoid) is used in this project to convert the outcome into categorical value to solve this problem. Thus, it is also one of the challenges in this project to understand the graph of logistic regression.

Moreover, the other challenge found in this project is the overfitting model. Train accuracy is the accuracy of a model on examples it was constructed on and the test accuracy is the accuracy of a model on examples it hasn’t seen. Overfitting model could be told from the large difference in accuracy between the test and train accuracy. For example, 45% of the train accuracy and 83% of the test accuracy were obtained in the project based on a model. But test accuracy should not be higher than train accuracy since the model is optimized for the latter. The validation and training accuracy should increase and loss should decrease when the final data had been checked which was used for training and data had been properly pre-processed. If some steps have gone wrong when data is being trained and tested, an overfitted model will occur.

CONCLUSION

With using machine learning techniques, three supervised learning models was built to implement in financial management field to ease the loan approval process. Three models which were Decision Tree Model, Random Forest Model and Logistic Regression Model were built by 6 researchers and each model handle by 2 researchers to accomplish this project. In order to make the prediction on the eligibility of a customer to get a loan more accurate, these 3 models were evaluated and the most suitable model was choosing for this project.

This project was beginning by choosing a suitable data set for the training and testing the models. After choosing the data set, there were some missing values in the data set and this will cause many effects like bias and inaccurate in the model. To prevent this problem, mice function was used to filled the missing parts. After the data set was clean, data visualization was carried out to determine the suitable variables to use in the models. This had made the models become more accurate in predicting.

From the result of building the three models, the decision tree model had a middle position of accuracy between 3 models. The final accuracy of decision tree model is 82.7%. For the last position of accuracy, the final accuracy is 78.92% which was from the logistic regression model. Lastly, the highest final accuracy was 84.86% obtained by the random forest model. The random forest model had the highest accuracy in predicting the eligibility of customer for loan application. Therefore, the most suitable model for this project is random forest model.

In conclusion, the loan approval prediction system was successfully built by three supervised learning model, Decision Tree Model, Random Forest Model, and Logistic Regression Model. These models were performing well in the prediction with three high accuracy results. But this system is not yet perfect, if this project has been given an extended period of time, the prediction of loan approval can include more significant features and increase the size of the data set to make the prediction more accurate.

FUTURE WORK

Financial management system is more and more important in this full of information and technology generation. This loan approval system can be improved and exploited to help the people in managing finances. Therefore, the researcher team of this project have gievn some suggestions to do the further research, if the given of time for this project has been extended.

The suggestion to improve the system is exploring the other model that can be used to predict the loan approval. K-Nearest Neighbor model and Support Vector Machine model can be considered because K-NN model is a classifier that very easy to implement and SVM model can performs very well when there is a clear margin of separation between classes.

Besides, exploiting the other factors of loan approval like current debt amount and current total assets, this might improve the accuracy of the proposed model in the future. The next suggestion is including more usable data to improve the quality of accuracy. A large number of observations with less missing values will train the model in a high-quality situation. This will makes the model more accurate when predicting the new data.

REFERENCES

1. Julia Kagan. (2021, Apr 19). Loan. Retrieved from https://www.investopedia.com/terms/l/loan.asp.

2. A. Vaidya, "Predictive and probabilistic approach using logistic regression: Application to prediction of loan approval," 2017 8th International Conference on Computing, Communication and Networking Technologies (ICCCNT), 2017, pp. 1-6.

3. Z. Ereiz, "Predicting Default Loans Using Machine Learning (OptiML)," 2019 27th Telecommunications Forum (TELFOR), 2019, pp. 1-4.

4. G. Arutjothi and C. Senthamarai, "Prediction of loan status in commercial bank using machine learning classifier," 2017 International Conference on Intelligent Sustainable Systems (ICISS), 2017, pp. 416-419.

5. Kumar Arun, Garg Ishan and Kaur Sanmeet, “Loan Approval Prediction based on Machine Learning Approach”. IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p-ISSN: 2278-8727,(May-Jun. 2016), PP. 18-21.

6. M. A. Sheikh, A. K. Goel and T. Kumar, "An Approach for Prediction of Loan Approval using Machine Learning Algorithm," 2020 International Conference on Electronics and Sustainable Communication Systems (ICESC), 2020, pp. 490-494.

7. Mitchell, T. M. (1999). Machine learning and data mining. *Communications of the ACM*, *42*(11), 30-36.

8. Tong, E. N., Mues, C., & Thomas, L. C. (2012). Mixture cure models in credit scoring: If and when borrowers default. *European Journal of Operational Research*, *218*(1), 132-139.

9. Granström, D., & Abrahamsson, J. (2019). Loan default prediction using supervised machine learning algorithms.

10. Arun, K., Sanmeet, K., & Ishan, G. (2014). Loan Approval Prediction based on Machine Learning Approach. *IOSR Journal of Computer Engineering*, *16*(3), 18–21. https://doi.org/10.9790/0661-1639 http://cloudstechnologies.in/cloudtech-admin/basepaperfiles/1593149297loan%20approval%20prediction%20using%20decision%20tree%20in%20python.pdf

11. Ashlesha Vaidya, Computer Science Engineering, SRM University, Chennai https://ieeexplore.ieee.org/document/8203946

12. Zhu, L., Qiu, D., Ergu, D., Ying, C., & Liu, K. (2019). A study on predicting loan default based on the random forest algorithm. *Procedia Computer Science*, *162*, 503-513. https://reader.elsevier.com/reader/sd/pii/S1877050919320277?token=243D6F2511200A867356E5120A5492CC584527BF69A159290BFCEEEBCB540B5A1D6057DBE190A739F50BA4C248E23353&originRegion=eu-west-1&originCreation=20210526053716

13. Mong'are, Dominic & Njoroge, Gladys & Muraya, Moses. (2019). Analysis of Individual Loan Defaults Using Logit under Supervised Machine Learning Approach. Asian Journal of Probability and Statistics. 3. 1-12. 10.9734/ajpas/2019/v3i430100. https://www.researchgate.net/publication/332785963_Analysis_of_Individual_Loan_Defaults_Using_Logit_under_Supervised_Machine_Learning_Approach

14. Sivasree M.S., & Rekha Sunny T. (2015). Loan Credibility Prediction System Based on Decision Tree Algorithm. https://www.ijert.org/research/loan-credibility-prediction-system-based-on-decision-tree-algorithm-IJERTV4IS090708.pdf

15. J. Tejaswini,T. Mohana Kavya, R. Devi Naga Ramya, P. Sai Triveni, Venkata Rao Maddumala(2020). Accurate Loan Approval Prediction Based On Machine Learning Approach. Journal of Engineering Science, 11(4), 523-532. https://jespublication.com/upload/2020-110471.pdf

16. ResearchOptimus. (2020, July 19). Logistic Regression and Its Application in Predicting Dependent Variables. Logistic Regression: Predicting Dependent Variables. https://www.researchoptimus.com/article/what-is-logistic-regression.php

17. Madaan, M., Kumar, A., Keshri, C., Jain, R., & Nagrath, P. (2021). Loan default prediction using decision trees and random forest: A comparative study. IOP Conference Series: Materials Science and Engineering, 1022, 012042. https://doi.org/10.1088/1757-899x/1022/1/012042

18. Why Data Visualization Is Important. Analytiks. (2020, June 10). https://analytiks.co/importance-of-data-visualization/#:~:text=Data%20visualization%20gives%20us%20a,outliers%20within%20large%20data%20sets.

19. What is a Decision Tree Diagram. Lucidchart. (n.d.). https://www.lucidchart.com/pages/decision-tree/#section_4.

20. Martinez-Taboada, Fernando; Redondo, Jose Ignacio (2020): Variable importance plot (mean decrease accuracy and mean decrease Gini).. PLOS ONE. Figure. https://doi.org/10.1371/journal.pone.0230799.g002

21. Pandey, P. (2020, July 29). Data Preprocessing : Concepts. Medium. https://towardsdatascience.com/data-preprocessing-concepts-fa946d11c825.

22. Brush, K., & Burns, E. (2020, February 20). What is data visualization and why is it important? SearchBusinessAnalytics. https://searchbusinessanalytics.techtarget.com/definition/data-visualization.

23. Alice, M. (2018, May 14). Imputing Missing Data with R; MICE package. DataScience+. https://datascienceplus.com/imputing-missing-data-with-r-mice-package/.

24. Kshitiz Gautam, Arun Pratap Singh, Keshav Tyagi, Mr. Suresh Kumar (2020). Loan Prediction using Decition Tree and Random Forest. International Research Journal of Engineering and Technology. https://www.irjet.net/archives/V7/i8/IRJET-V7I8145.pdf

25. Mong'are, Dominic & Njoroge, Gladys & Muraya, Moses. (2019). Analysis of Individual Loan Defaults Using Logit under Supervised Machine Learning Approach. Asian Journal of Probability and Statistics. 3. 1-12. 10.9734/ajpas/2019/v3i430100. https://www.researchgate.net/publication/332785963_Analysis_of_Individual_Loan_Defaults_Using_Logit_under_Supervised_Machine_Learning_Approach

26. Zhu, Lin & Qiu, Dafeng & Ergu, Daji & Ying, Cai & Liu, Kuiyi. (2019). A study on predicting loan default based on the random forest algorithm. Procedia Computer Science. 162. 503-513. 10.1016/j.procs.2019.12.017. https://www.researchgate.net/publication/338286615_A_study_on_predicting_loan_default_based_on_the_random_forest_algorithm