Practical Business Analytics Report

Lending Club 2020

Group Name: 6Tech

Surrey Email	Member Name	University ID
ms02719@surrey.ac.uk	Muaad Siala	6664948
gg00498@surrey.ac.uk	Gowrisaranyan Ganesan	6654496
ky00241@surrey.ac.uk	Krishna Kanth Yarraboina	6657929

Table of Contents

Section 1: Problem Definition and Data Understanding

1.1 Problem Definition:

The Lending Club is a US peer-to-peer lending company with its headquarters located in San Francisco, California. It was the first peer-to-peer lender to register its offerings as tradable financial assets with the U.S Securities and Exchange Commission, and to offer secondary loan trading. Lending Club is the world's peer-to-peer lending platform.

It enables borrowers to create unsecured personal loans between $1,000 and $40,000 and the standard loan period is years. Investors are provided the options to search and select their desired loan among others listed on the lending club website based on the information given about the borrower, the amount of loan, loan grade and loan purpose.

These loans thereby generate interest borrowers and origination fees for their service. It is transforming the banking system to make credit more affordable and investing more rewarding. Lending Club operates at a lower cost than traditional bank lending programs and passes the savings on to borrowers in the form of lower rates and to investors in the form of

Although each borrower's background is subjected to a thorough examination before their loans are permitted, there is a moderate proportion of the loans that have not been repaid by the borrowers they are incapable of financially.

This project aims to investigate the dataset where the loans have already been accepted to assess whether the potential borrower candidate's loan will be charged off. This can help potential investors to make a sound decision about investing. A machine learning model that can predict this probability will help investors make better decisions.

Objective: To build a machine-learning model that can take into consideration important features like FICO scores, debt to income ratio about investing and predict the probability of the borrower candidate's loan being charged off.

1.2 Business Analytics Tasks:

In this project, we are following the Cross Industry Process for Data Mining (CRISP-DM) approach. The CRISP-DM approach contains six essential steps, which are Business Understanding, Data Understanding, Data Preparation (Pre-processing), Modelling, Evaluation, Deployments. See **figure 1**.

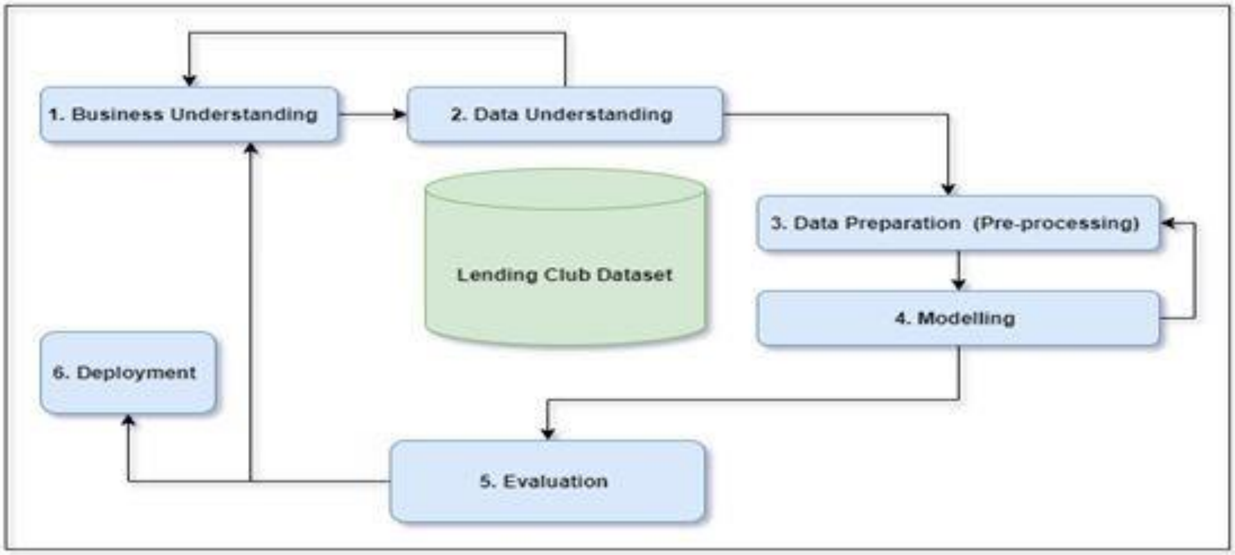


Figure 1 shows the Cross-Industry Process for Data Analytics.

1.3 Data Understanding:

In data understanding we divide the description of the data into four parts, the first part is The Initial Data collection, the second part is Describe Data, the third part is Explore Data, and fourth part is Verifying the Data Quality.

1.3.1 Initial Data Collection:

In this project, we are going to use and analyse the Lending Club dataset from Kaggle. The dataset is divided into two main parts. The first part is the loans that were accepted by the lending club from 2007 to 2018, The second part is the loans that were rejected by the lending club from 2007 to 2018. In this project, we will dive into the accepted loans as shown in **figure 2.**

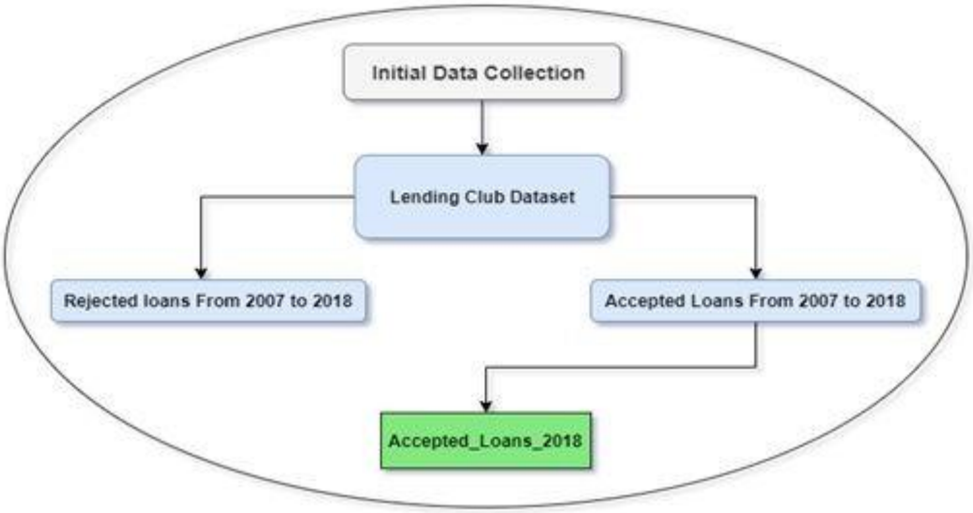

Initial Data Collection
Lending Club Dataset
Rejected loans From 2007 to 2018
Accepted Loans From 2007 to 2018
Accepted_Loans_2018

Figure 2 showing the narrow down of the dataset collection.

1.3.2 Data Exploration:

The dimensions of the accepted loans 2018 dataset are the number of columns are 151, number of rows 32834. In this project, we selected this dataset just to make sure our dataset is suitable to answer the following two questions:

1) Would the borrowing person get charged – off?

2) Did this person end up paying back their loan?

After selecting relevant features, selecting relevant records, and getting necessary data details. We explored the number of categorical features in the dataset is 35, and the number of numerical features in the dataset is 113. We have 3 logical columns as well.

1.3.3 Data Dictionary:

There are 122 features in the dataset and the dataset is mixed with numerical and categorical variables.

Also it includes integers, characters, double, string and in the following table 1 we are mentioning the most important columns such as our target variable is loan_status.

The entire table for the data dictionary can be found in the appendix section.

Column name	Description
loan_status	Loan status contains two values, the first one is when the borrowers fully paid and the second is when the borrowers is charged off.
dti	A ratio calculated using the borrower's total monthly debt payments on the total debt obligations, excluding mortgage and the requested LC loan, divided by the borrower's self-reported monthly income.
loanAmnt	The listed amount of the loan applied for by the borrower. If at some point in time, the credit department reduces the loan amount, then it will be reflected in this value.
zip_code	The first 3 numbers of the zip code provided by the borrower in the loan application.
purpose	A category provided by the borrower for the loan request.
ficoRangeHigh	The upper boundary range the borrower's FICO at loan origination belongs to.
ficoRangeLow	The lower boundary range the borrower's FICO at loan origination belongs to.
grade	LC assigned loan grade
subGrade	LC assigned loan subgrade
application_type	Indicates whether the loan is an individual application or a joint application with two co-borrowers

Table 1 contains ten of the most important columns in the dataset.

1.3.4 Verification of Data Quality:

We are certain that the data meets the requirements of our purpose of use in terms of Accuracy, Relevancy, Completeness, Timelines, and Consistency. The ‘Accepted Loans 2018’ dataset has been verified by Kaggle.com and it is already used in other similar projects.

1.3.5 Identifying the Target Variable:

In the Accepted Loans dataset, we have identified the target variable as "Loan Status". The Loan Status has 10 categories. For our business objective, we are considering fully paid and charged off class under "Loan Status". Loans that are current, do not meet the credit policy, defaulted, or have a missing status will not be considered. We plan to focus on results from the year 2018.

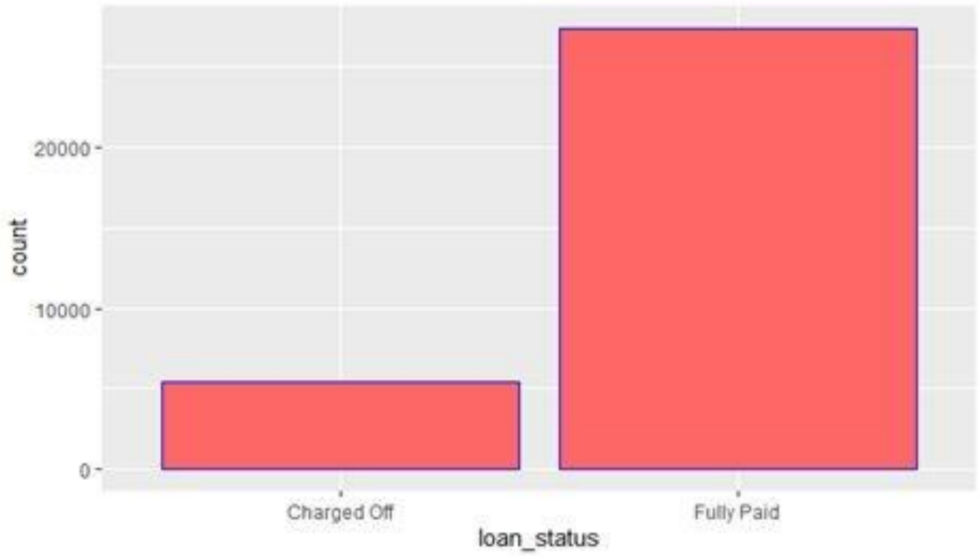


Figure 3 shows the distribution of the loan status.

Figure 3 is a graph showing distribution of the target variables charged off and fully paid. We can see the uneven distribution of two classes in the target variable. Percentage of Fully Paid borrowers is 16.4 whereas percentage of charged off borrowers is 83.6.

1.3.6 Data and libraries loading:

The dataset is loaded into the R environment, simply by using read.csv function with path file "Accepted_Loans_2018”.We imported the libraries required for data pre-processing and data loading. The following libraries like Lattice, Caret, Repr, Tidy verse, Dplyr are used in various preprocessing stages.

Section 2: Data Preparation (Pre-processing)

2 Data Pre-processing:

In the data preparation section, we will be going into four parts of data pre-processing. The first part is the Data and Libraries loading. The second part is Data Exploring and Cleaning. The third part is Finding Correlation between variables. The fourth part is Outliers Removal.

2.1 Data Exploring and cleaning:

- We have 32834 rows and 151 columns.
- Number of categorical features are 35 whereas numerical features are 113.
- We are keeping columns which have null values less than 30%.
- Number of columns we are left with after removing columns from the dataset which have more than 30% of null values are 119.
- Most of the columns in the dataset have less than 5 percent of null values.
- Ongoing through Data Dictionary, we found the features in table 2 are to be relevant for data analysis.

addr_state	***annual_inc***	***application_type***	***Dti***
earliest_cr_line	***emp_length***	***emp_title***	***fico_range_high***
fico_range_low	***grade***	***home_ownership***	***Id***
initial_list_status	***installment***	***int_rate***	***issue_d***
loan_amnt	***loan_status***	***open_acc***	***mort_***
pub_rec_bank	***pub_rec***	***Purpose***	***zip_code***
revol_util	***sub_grade***	***Term***	***Title***
verification_status	***total_acc***	***revol_bal***	**-**

· **Table 2** Shows the most relevant features that fit our models.

- FICO Scores can be mathematically computed by applying the formula on fico range low and fico range high.

```
FICO Score = 0.5 * Fico Range Low + 0.5 * Fico Range Low
```

(a) Removal of further Categorical columns:

Even though the features in **table 2** are useful for pre-processing, not all columns are needed for Modelling.

After analysing the data dictionary, we performed chi-square analysis on the certain pairs of the input categorical columns to determine the association (correlation) between them.

We removed columns as we found strong evidence in chi-square test these columns have strong association between them. We removed one column from each pair. Keeping them will only affect the model while prediction of target variable.

On performing chi square analysis we found that:

- Grade column is found to have a strong correlation with subgrade. As the grade is implied by subgrade so we removed them.

- Zip code is considered a categorical column as the number in Zip-code describes a unique location. Zip-code is found to have strong correlation with Address state, so we removed the Zip code column.

- Purpose and Title are found to have strong correlation. We removed the title column. Purpose and title mentions the purpose of borrowing the loan.

See figure 4 below, a graph shows the top 5 percentage of Borrower's title for purchasing a loan from investors. We can clearly see that many used their loan debt consolidation.

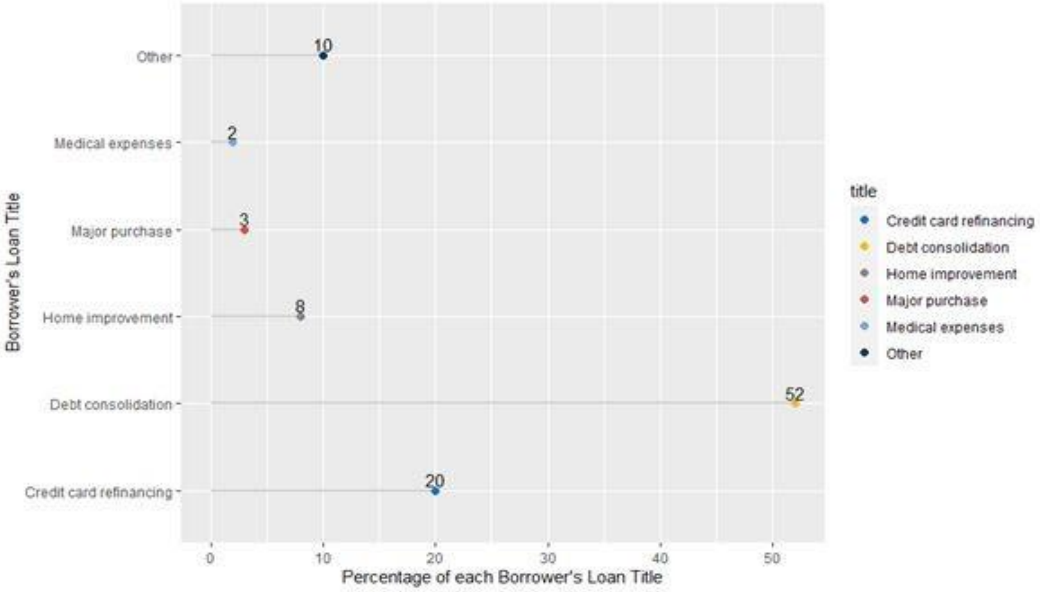


Figure 4 shows the top 5 percent of a borrower's title for purchasing a loan.

(b) Removal of Null Values:

There are blanks in the dataset which did not detect them as N/A or null values. So, all the rows wherever empty spaces are detected are dropped from the dataset.

(c) Manipulating the values of columns:

- emp_length column mentions years of experience of each borrower. We changed the people who are having less than one year of experience as 0 year and then 10 + years as 11 years of experience. We removed the strings from this column and changed it to numeric type.
- We performed the above steps on term count and converted it into numeric type.
- On earliest_cr_line, It is in the pattern of month and year (Month-YY). We modified the column by removing month and retaining only year. That way it is converted into numeric type.

(d) Removing EMP_TITLE column:

emp_title describes the designation of each borrower. This column is removed from the dataset as the emp_title has more unique categories in it. There are about 14,349 categories in the emp_title column.

(e) Checking the Subcategories in categorical columns:

We reduced categorical columns from 35 to 9. On checking the subcategories we found that many categorical features have less than 6 subcategories. We are keeping sub-grade and addr_state columns, even though they have more than 30 distinct subcategories.

2.2 Correlation between Numerical variables:

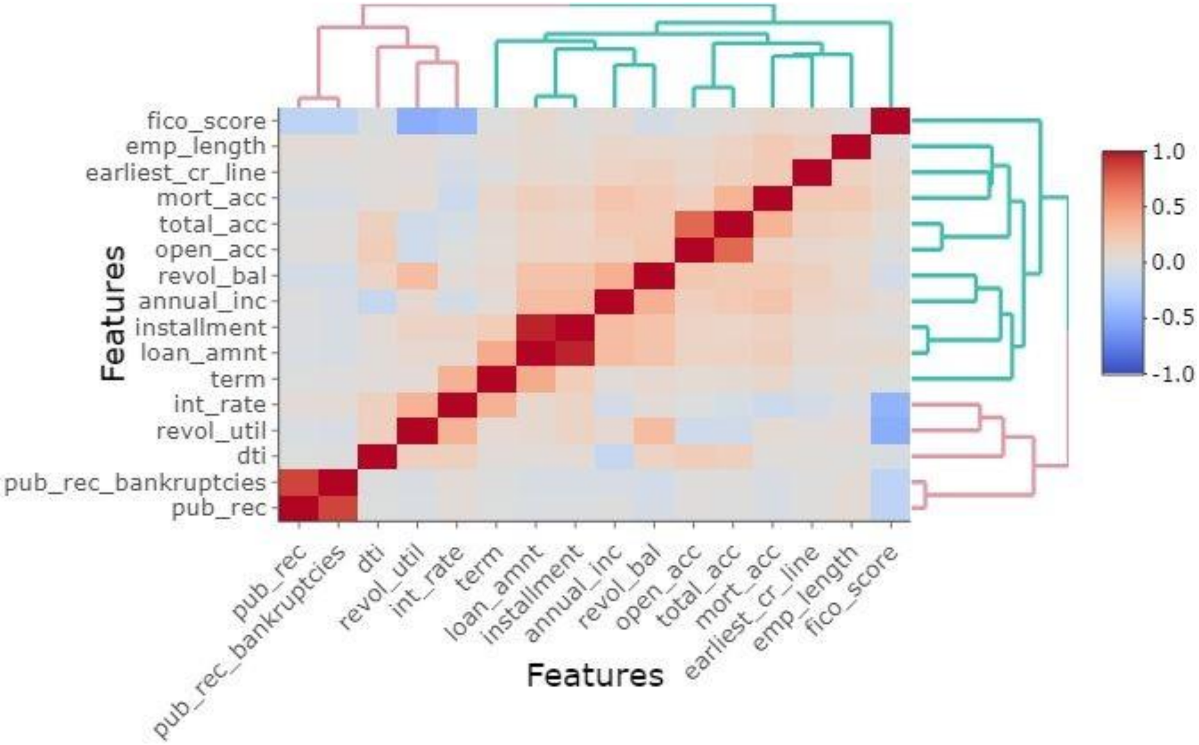


Figure 5 heatmap plot shows the correlation among numerical variables.

Input Numerical variables have strong correlation between them. Keeping them in the dataset will affect our analysis. It create problems during modelling like overfitting and extra dimensionality .After applying the correlation function on numerical columns, we removed:

- · loan_amnt column as it has strong association with installment.

- · open_acc column as it has strong association with total_acc.
- · pub_rec_bankruptcies as it has strong association with pub_rec.

2.3 Removal of Outliers:

We checked outliers in every column and removed only extreme outliers from certain columns. We followed the below steps:

- Visualized the box plot on each numerical column.
- Removed the extreme outliers by specifying the particular percentile. For example, we removed records from annual income values by removing the values which are above 99 percentile.
- Visualized the boxplot after removing the extreme outliers. (See figures 6 & 7)

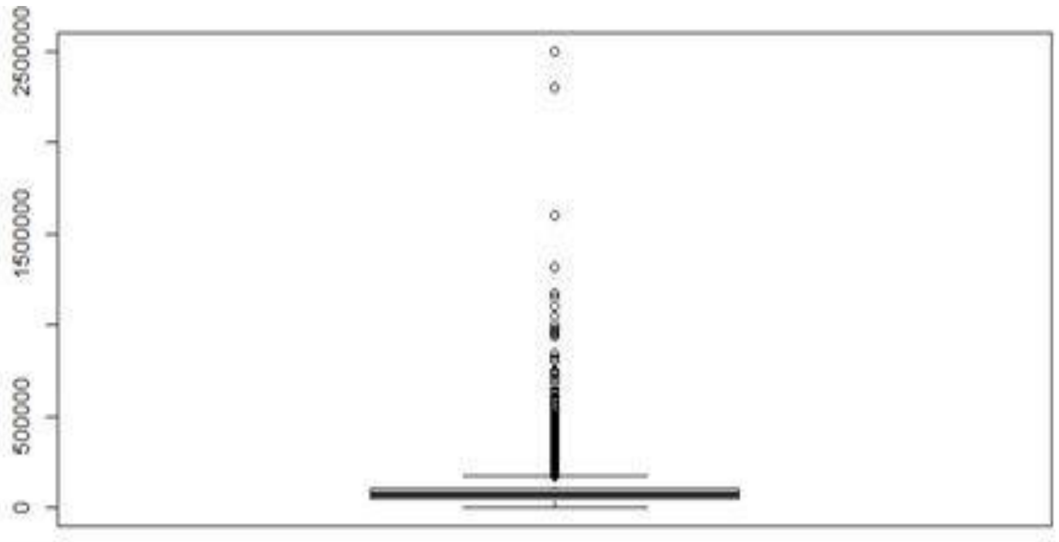


Figure 6 visualization of box plot on annual income column before removing extreme outliers.

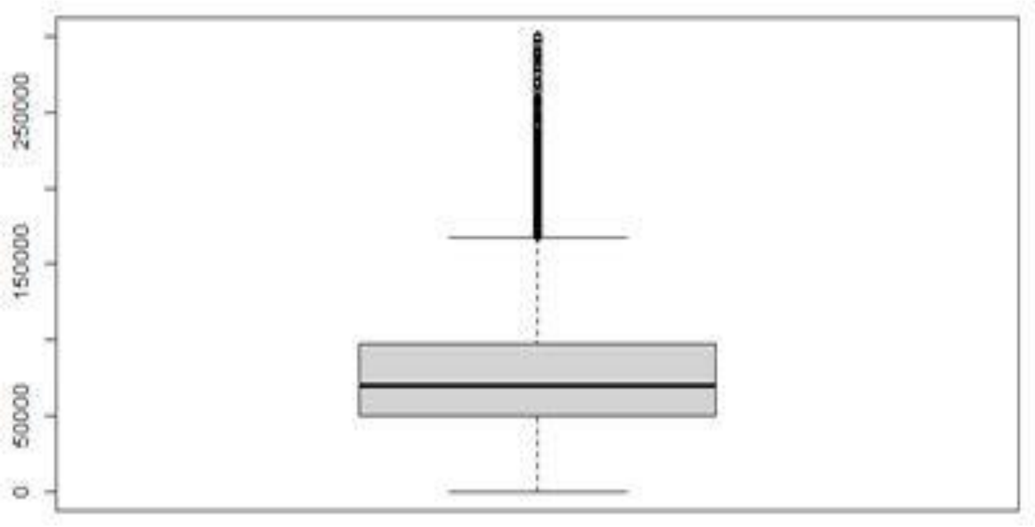


Figure 7 visualization of box plot on annual income column after removing extreme outliers.

2.4 Data Visualization:

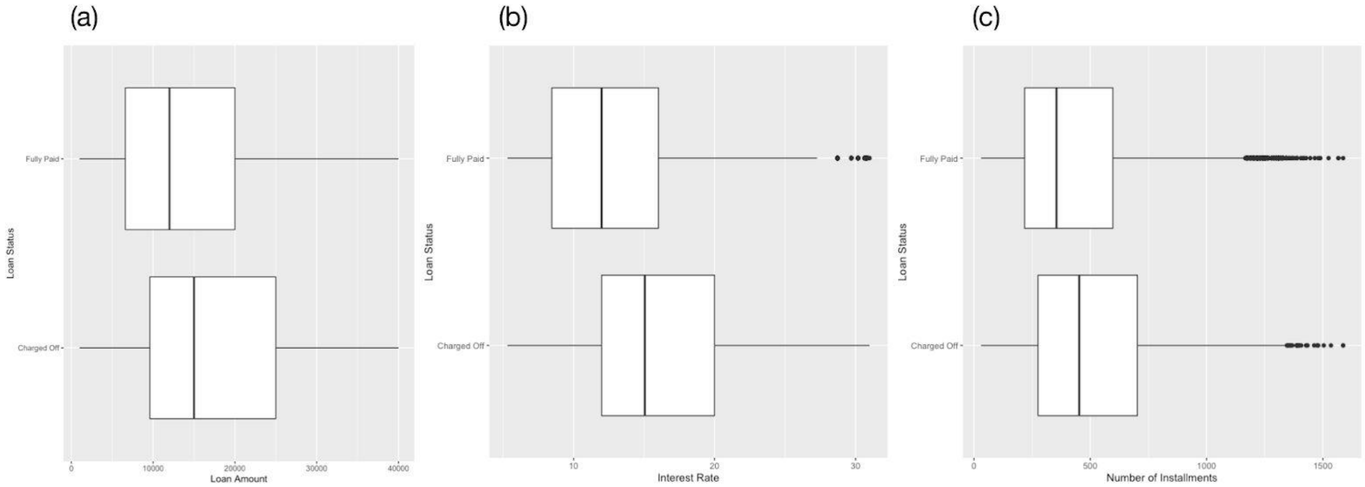
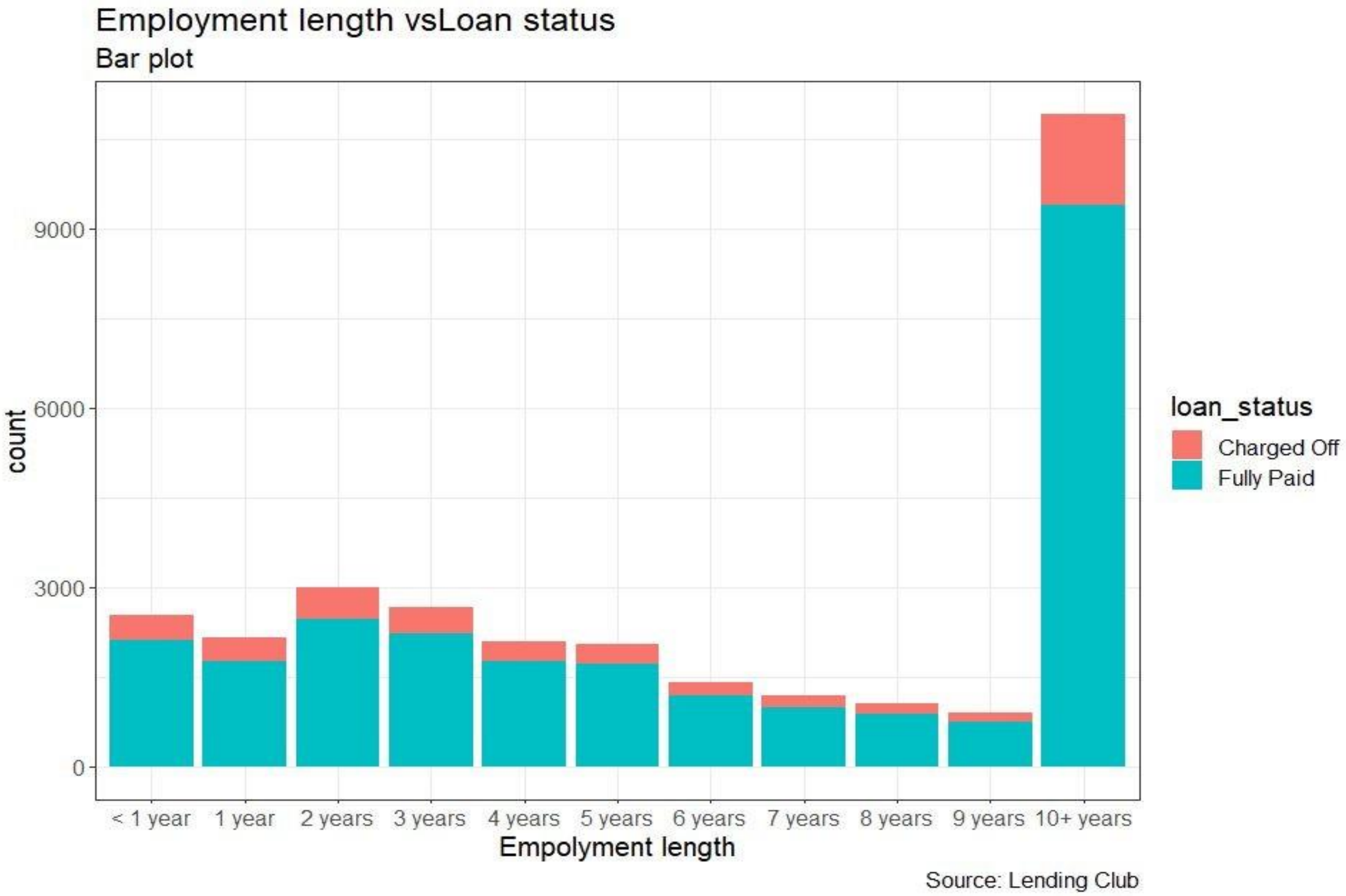


Figure 8. The three graphs represent the relationship of loan status against 3 attributes: a) loan amount, b) interest rate and c) number of instalments respectively. The box plots clearly highlight the significant differences in the three categories amongst the Fully Paid loans and Charged Off loans. It is evident in Figure 4 that higher loans, interest rates and number of installments are associated with higher number of charged off loans.



Source: Lending Club

Figure 9. demonstrates a bar chart that plots the distribution of employment against loan status.

This chart illustrates the relationship between employment length and loan status.It is apparent that borrowers who have worked more than 10 years are more likely to have their loan accepted therefore they have the highest fully paid loans count and charged off loans.

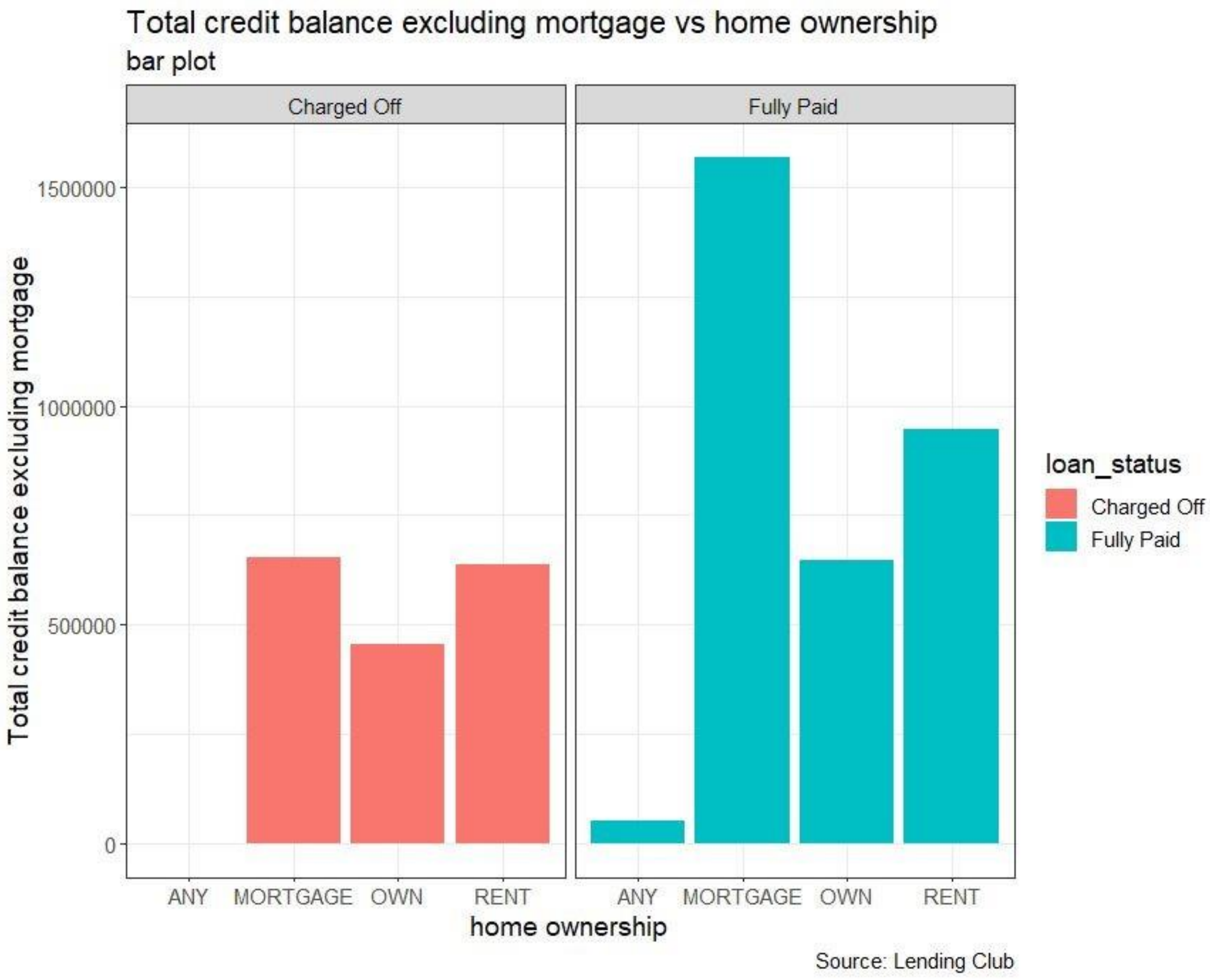


Figure 10. shows bar chart investigates the relationship between total revolving balance excluding mortgage and home ownership.This chart is split into charged off loans and fully paid loans.In the fully paid loan section,borrowers who is living in a mortgaged property has the highest number which are the group who is the most likely to pay back the loans.It is then loosely followed by renters and owners.

In the charged off loans section,the number of charged off loans from borrowers who live in mortgaged property are closely followed by the renters.This however does not directly indicate borrowers who live in mortgaged property are the more likely to not pay.It is clear that property owners are more likely to not pay back the loans as the number of charged off loans and paid back loans are very close which imply the chances of owners are almost halved.

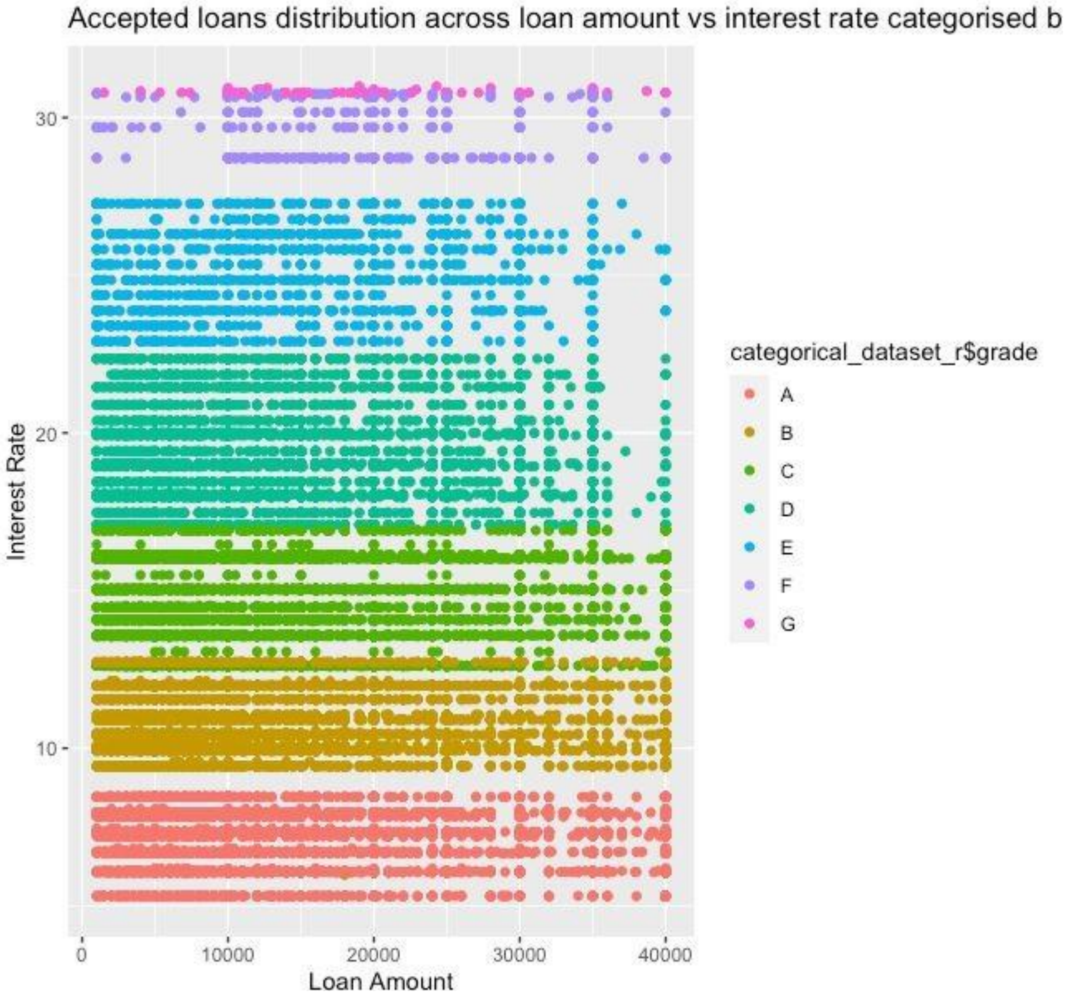


Figure 11. highlights the interest rate offered on loan amounts to customers with different grades. The scatter plot depicts that interests rates significantly increase for customers as we proceed from grade A to grade G, for the same loan amount. This shows that interest rate is not dependent on the loan amount, but on the grade assigned to the customer. This, however, does not impact the number of loans granted to grade A, B,C,D and E. However, the number of loans granted are extremely low for grade F and G possibly due to higher interest rates.

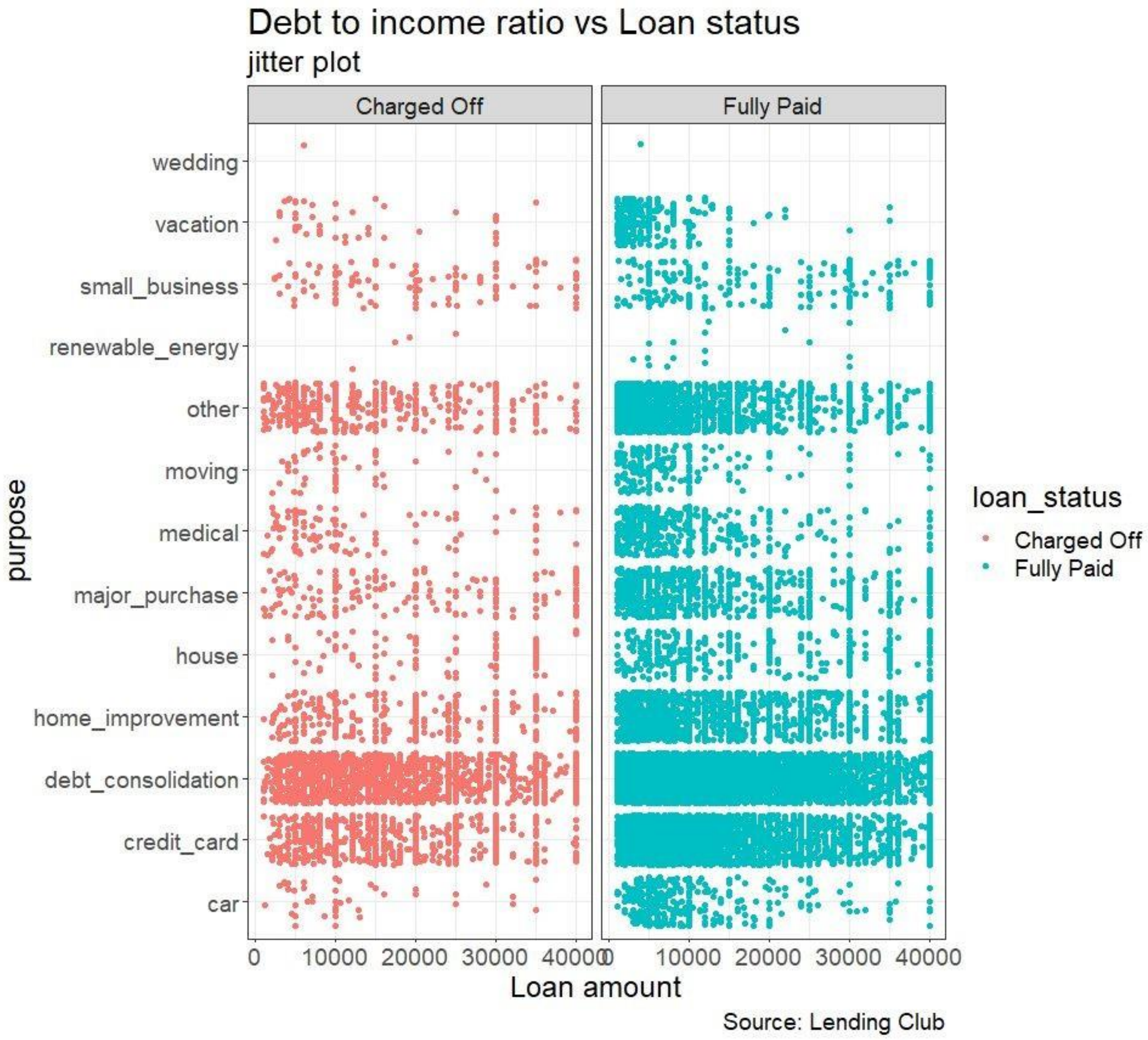


Figure 12. plots the relationship between purpose and loan amount

This graph from lending club shows purposes that loans were taken out for and separate the loan with charged off status and fully paid status.From this charged off section we can see that loans taken out for credit card ,debt consolidation ,other purposes has the most dots out of all the purposes.Moreover we can see a trend where the number of dots increases as the loan amount decreases.This is possibly because the loan amount is a significant factor for lending club to assign a loan grade to a accepted loan.When loan amount for a loan reduces,it is easier for lending club to misjudge a borrower creditworthiness .

On the other hand,in the fully paid section we can also observe loans taken out for credit card ,debt consolidation ,other purposes has the most dots out of all the purposes.This means that a lot of people borrow money from lending club when they have financial difficulty with credit card ,debt consolidation ,other purposes.

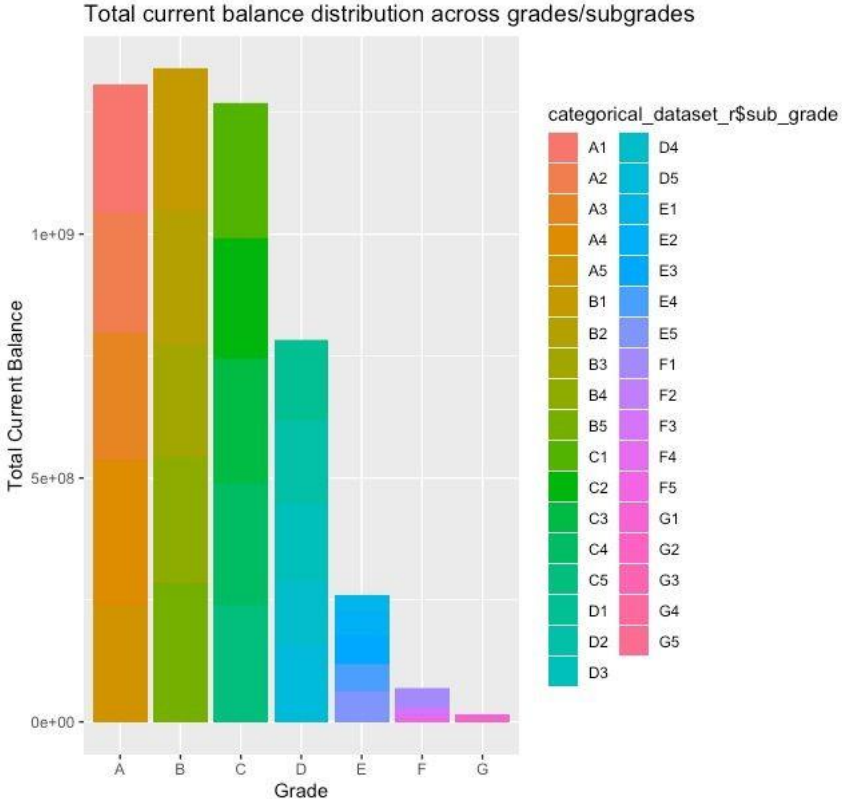


Figure 13. Total current balance distribution across different grades. The bar chart demonstrates that grades A, B and C have almost similar total current balance in their account. There is a steep decrease in the balance from grade C to grade D, grade D to grade E, grade E to grade F and extremely low for grade G.

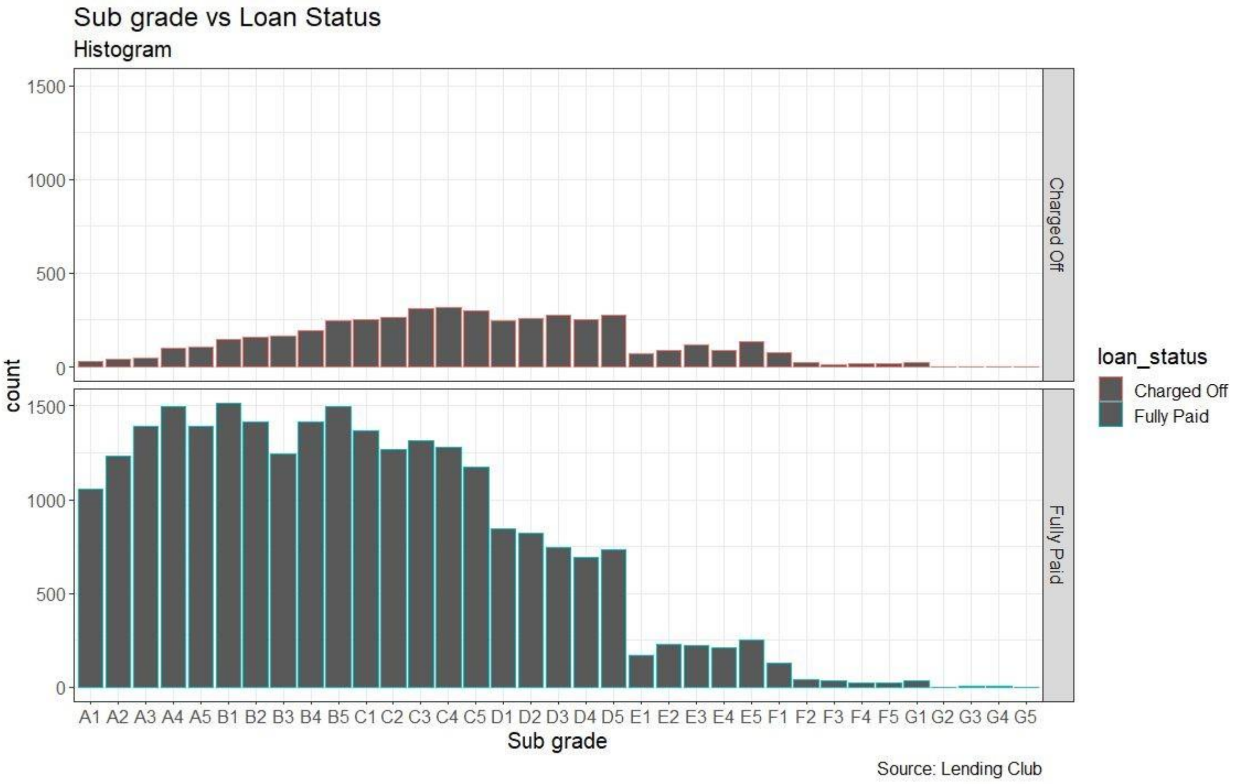


Figure 14 shows the distribution of Sub grade separated by loan status

Every loans is assigned a subgrade according to the borrower's information such as FICO score ,annual income ,loan amount etc.This .There are 7 grades from A to G and each grade is subdivided into 5 levels ranging from 1 to 5.A1 is the best possible grade given to the loan and G7 is the worst possible grade can be assigned to a loan.

This chart compares subgrade and loan status which gives us insights which subgrade is not assigned properly which lead to charged off loans.In the charged off loan section, we can see the numbers steadily increase from A1 to C4 in addition to that,number between B5 and D5 consistently flucatated above 250.Although between B5 and D5,C4 grade has the most charged off loans ,C4 also has one of the highest number in the fully paid section.

In the fully paid section,the number started dramatically decrease starting from C5 and down to 750 around 750 between D1 to D5.Furthermore, from Grade E to Grade G ,it has the most similar charged off loans number and fully paid number which means majority of the loans with grade E to G should not be accepted and should be rejected from the beginning.

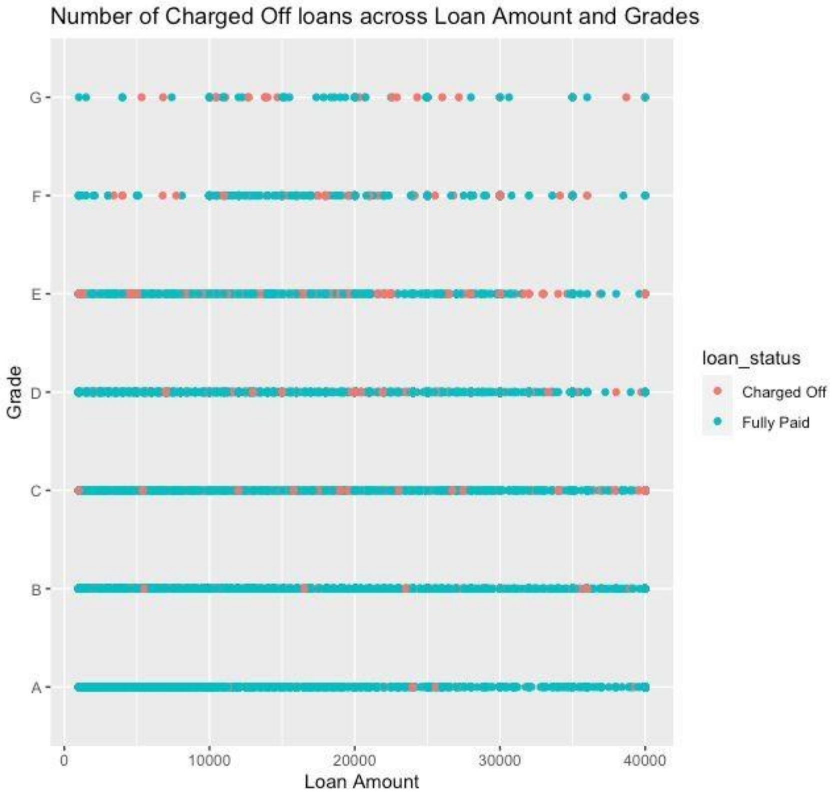


Figure 15. Represents the number of Charged Off loans spread across loan amount vs grade. The number of charged off loans increase as we go from grade A to grade G as shown by red values. This confirms that lower grades and higher interest rates together negatively affect loan repayment.

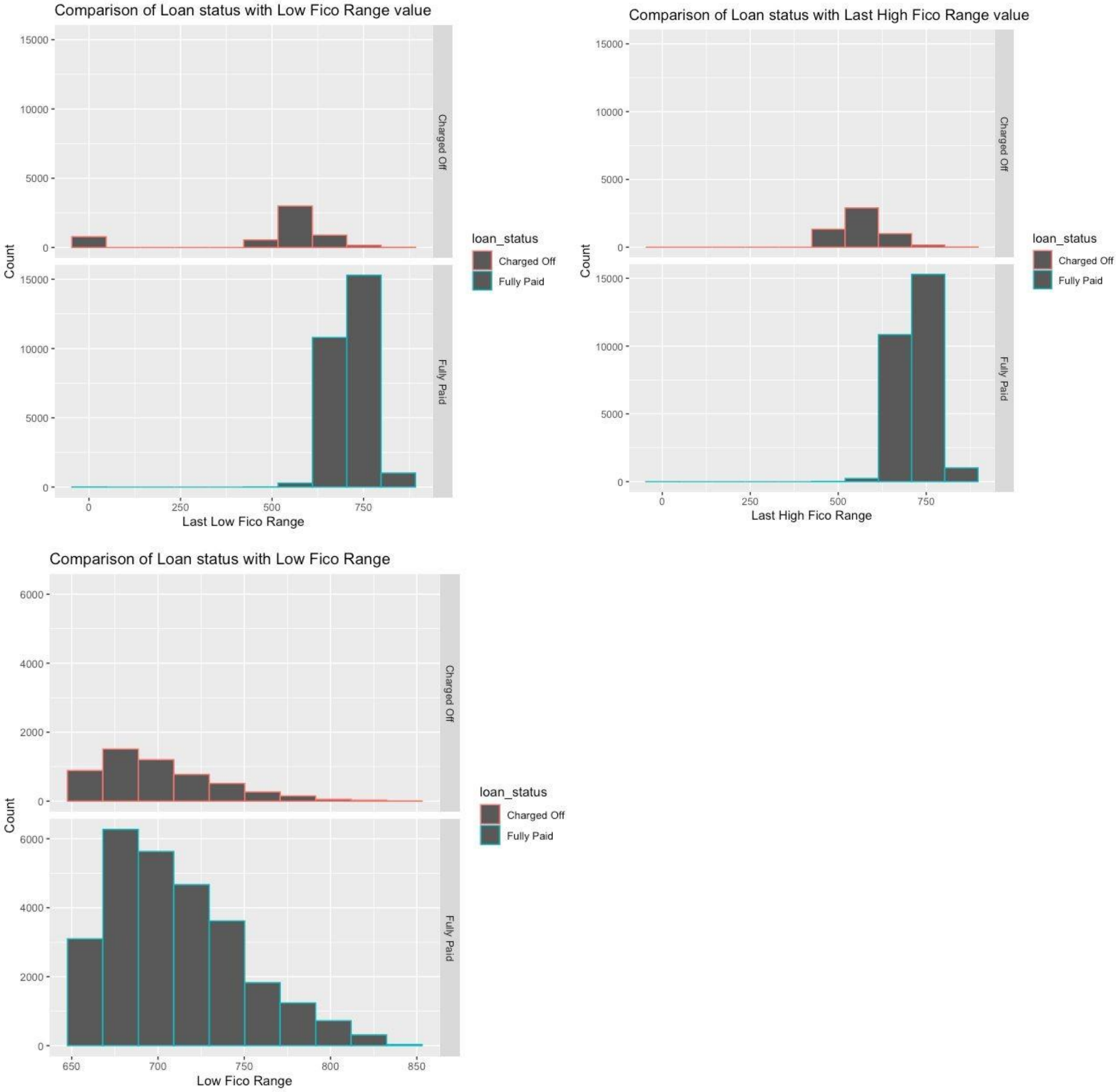


Figure 16. The histogram represents low Rico range values against loan status. a) Low Fico Range values do not provide any useful relationship between charged off and fully paid loans as the data is evenly distributed in ratios across both categories. b) The last Rico range values below approx. 640, however, shows an increase in the number charged off loans compared to fully paid loans, which is an important determinant that can be used as a deciding factor by investors. c) Comparing higher fico range value against loan status confirms the point we state in part b.

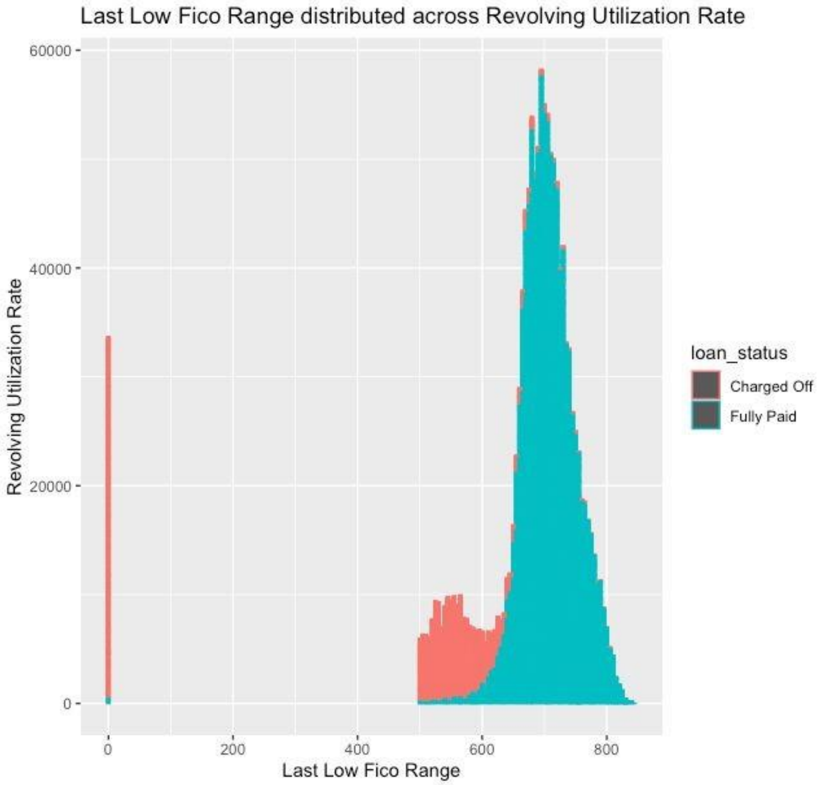
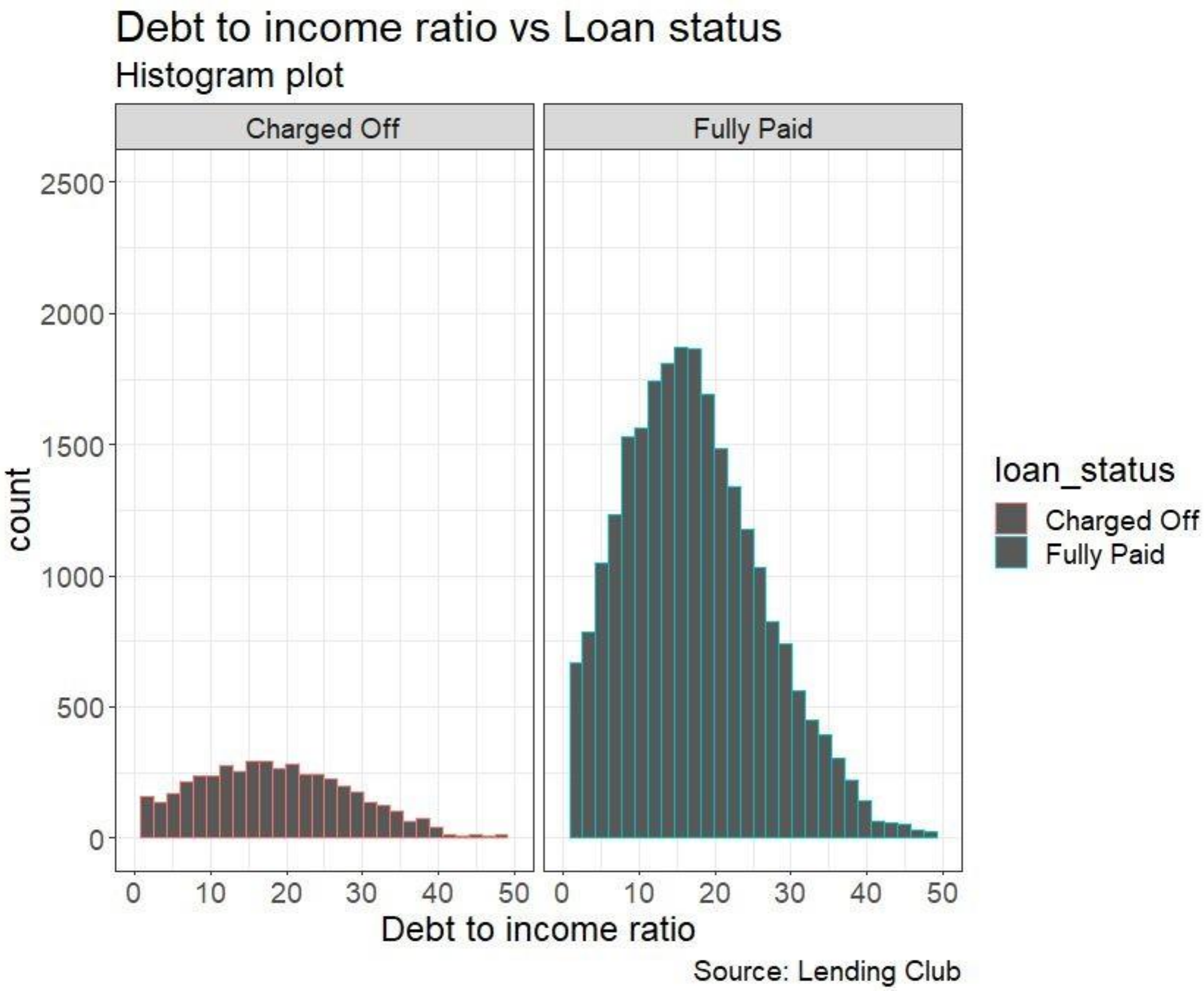


Figure 17. Last Fico Range v/s Revolving utilisation rate. The graph highlights a significant red zone which depicts the charged off loans. We can conclude from the graph that the last fico range value below approx. 620 combined with a revolving utilisation rate below 10000 can be used a factor by investors to prevent investing in risky loans.



Debt to income ratio vs Loan status

Histogram plot

Charged Off

Fully Paid

Source: Lending Club

Figure 18 shows debt to income ratio separated by loan status

This plot investigates the effect of debt to income ratio on loan status. Debt-to-income ratio(DTI) compares the total amount you owe every month to the total amount you earn.It is a industry standard evaluation factor to lenders in tandem with credit reports and FICO score when weighing credit applications.A DTI needs to be below 43 to be considered as good and still qualified for a loan.

From this plot, there is a general uptrend towards the middle range of the DTI and number plummets as it approaches to both ends of the chart .THis trend is seen in both charged off and fully paid loans.In addition, we can see that as DTI rises, the number of both status starts to synochoised and have a similar level.This means that Lending club should look out for the borrowers with high DTI as it is very likely that they will not be able to pay back the money.

Section 3: Modelling and Evaluation

A great part of machine learning is classification, we want to know what class an observation belongs to. The ability to precisely classify observations is extremely valuable for various business applications like predicting whether a particular user will buy a product or forecasting whether a given loan will default or not.

In this project, after the preprocessing stage we will feed the sub-set data into 6 models and compare the result between these models by using confusion matrix, precision, recall and f-measure.

3.1 Model Building:

After the preprocessing stage we reach to modelling our machine learning models. Before feeding the pre-processing stage into models, we will go into more pre-processing to fit our dataset into models. The more pre-processing stage includes 5 steps:

3.1.1 Splitting the data set:

The dataset has been split to train data and test. Train data consists of 70% of data and test data contains 30% of data. In test data, target Variable has been removed because we need to predict the target variable.

3.1.2 Evaluate the importance variables:

There are two measures of importance for every variable in the random forest. The first one is based on how much the accuracy decreases when the variable is removed. This is further broken down by outcome class. The second measure is based on the decrease of Gini impurity when a variable is chosen to split a node.

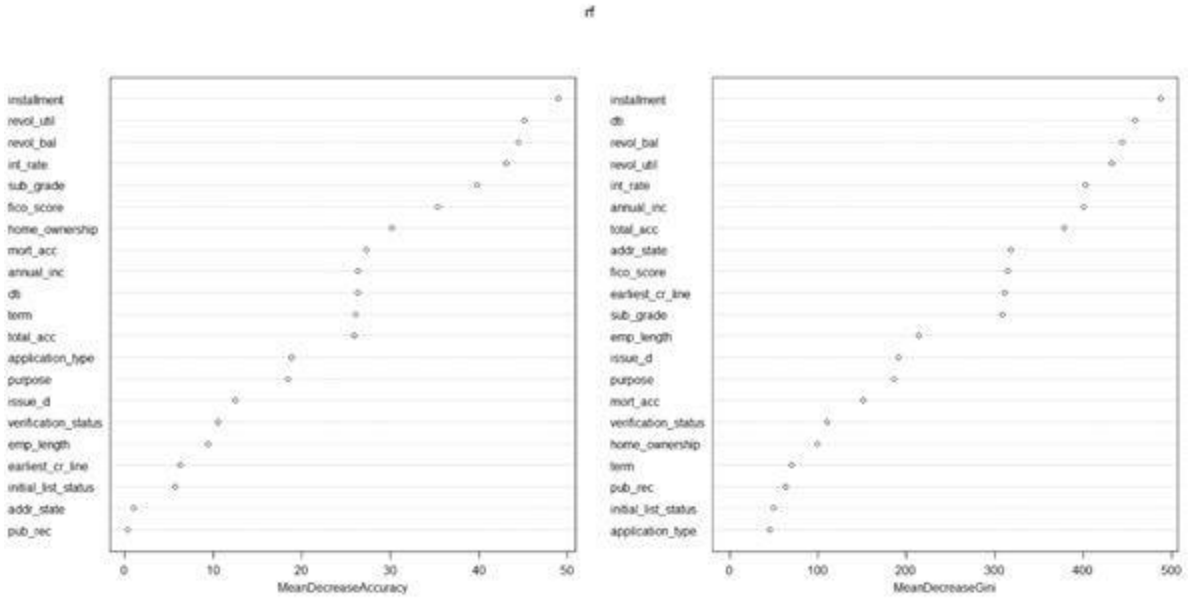


Figure 19 shows features importance evaluation.

Feature importance is a technique that assigns a score to input features based on how useful they are at predicting a target variable.

3.1.3 Encoding the categorical variables:

After getting the importance variable by the random forest then we proceeded to encode the categorical variables by assigning the dummy variables to the categorical variables of the preprocessed data.

- Dummy variables are commonly used in statistical analysis and in more simple descriptive statistics.
- A dummy column is one which has a value of one when a categorical event occurs and zero when a categorical event doesn’t occur.
- Then we want to assign the dummy variables to the dataframe.

3.1.4 Smote analysis for balancing the target variables:

- In this dataset the target variable is imbalanced, so we need to balance the target variable.
- For balancing the target variable we are performing smote analysis on the target variable in the dataset.
- After Smote analysis it's clearly seen that the target variable is balanced.

3.1.5 Split the dataset for the data modelling (After Encoding):

- We were splitting the dataset after encoding the dataset to perform the data modelling.

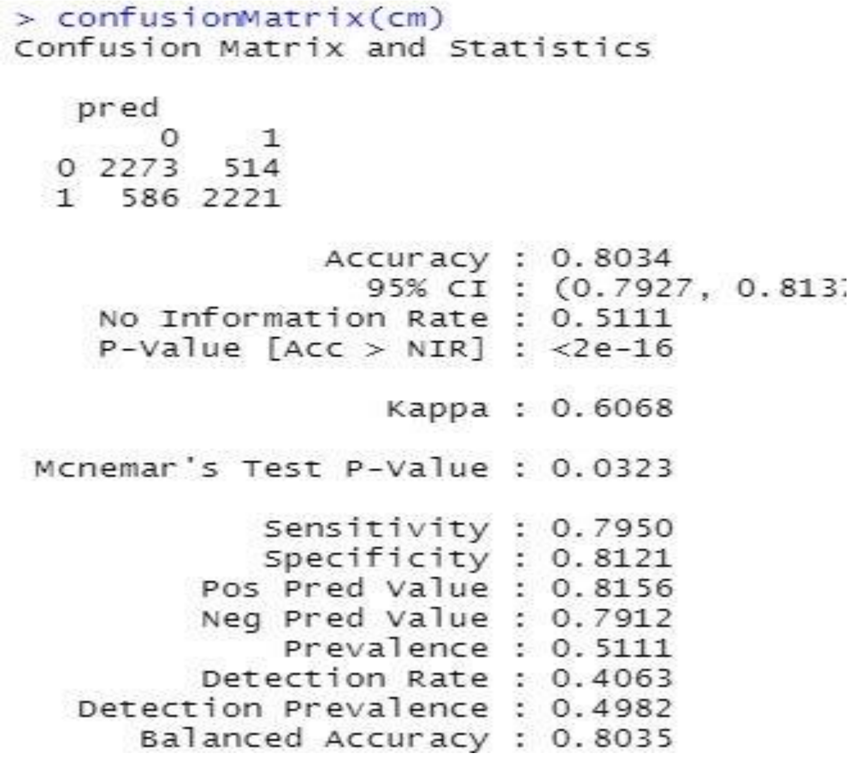
3.2 Random Forest Model:

- Random forest consists of a greater number of independent decision trees that operate as groups.
- Each individual decision tree in the random forest predicts out a class prediction and the class with the highest votes becomes our model's prediction.
- For random forest modelling, we are using random forest function to train the model with trained data.
- Then we predict the data with the trained data and to measure the performance of the algorithm we are using the confusion matrix metric.

3.2.1 Random Forest Result:

The Random Forest model achieves an accuracy up to 80.34%.

3.2.2 Random Forest Evaluation:



```
> confusionMatrix(cm)
Confusion Matrix and Statistics

     pred
         0    1
   0 2273  514
   1  586 2221

                Accuracy : 0.8034
                  95% CI : (0.7927, 0.813
     No Information Rate : 0.5111
     P-Value [Acc > NIR] : <2e-16

                   Kappa : 0.6068

  Mcnemar's Test P-Value : 0.0323

             Sensitivity : 0.7950
             Specificity : 0.8121
          Pos Pred Value : 0.8156
          Neg Pred Value : 0.7912
              Prevalence : 0.5111
          Detection Rate : 0.4063
    Detection Prevalence : 0.4982
       Balanced Accuracy : 0.8035
```

Figure 20 confusion matrix for Random Forest.

Classification Metrics:

True positive: data points labeled as positive that are actually positive

False positive: data points labeled as positive that are actually negative

True negatives : data points labeled as negative that are actually negative

False negative: data points labeled as negative that are actually positive

Recall and Precision Metrics:

Precision : ability of a classification model to return only relevant instances.

Precision = true positives/true positives+false positives

Recall: ability of a classification model to identify all relevant instances

Recall = true positives/true positives+false negative

F1 score: single metric that combines recall and precision using the harmonic mean

F1 score = 2 * precision*recall/precision+recall

Accuracy : Accuracy is one metric of evaluating classification models

Accuracy = number of correct predictions/total number of predictions

Precision of the random forest 80.41

Recall of the random forest 81.89

F1 score of the random forest 81.15

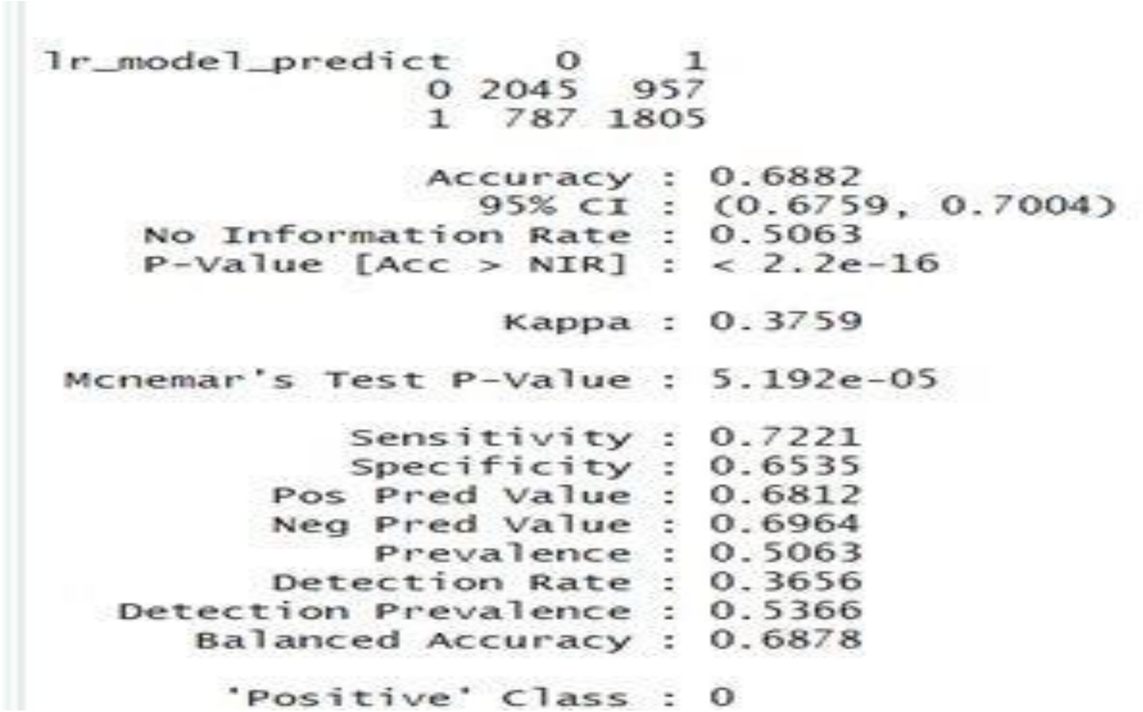
3.3 Logistic Regression:

- Logistic regression is used to predict the class of individuals based on one or multiple predictor variables.
- Logistic regression does not directly return the class of observations and it allows us to estimate the probability of class.
- For logistic regression model we are using glm function to train the training dataset.
- Then we predict the trained dataset module and to evaluate the model we are using the confusion matrix.

3.3.1 Logistic Regression Result:

The Logistic Regression model achieves an accuracy up to 68.82%.

3.3.2 Logistic Regression Evaluation:



```
lr_model_predict    0    1
               0 2045  957
               1  787 1805

               Accuracy : 0.6882
                 95% CI : (0.6759, 0.7004)
    No Information Rate : 0.5063
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3759

 Mcnemar's Test P-Value : 5.192e-05

            Sensitivity : 0.7221
            Specificity : 0.6535
         Pos Pred Value : 0.6812
         Neg Pred Value : 0.6964
             Prevalence : 0.5063
         Detection Rate : 0.3656
   Detection Prevalence : 0.5366
      Balanced Accuracy : 0.6878

       'Positive' Class : 0
```

Figure 21 confusion matrix for Logistic Regression.

Precision of the Logistic Regression 67.75

Recall of the Logistic Regression 70.43

F1 score of the Logistic Regression 69.18

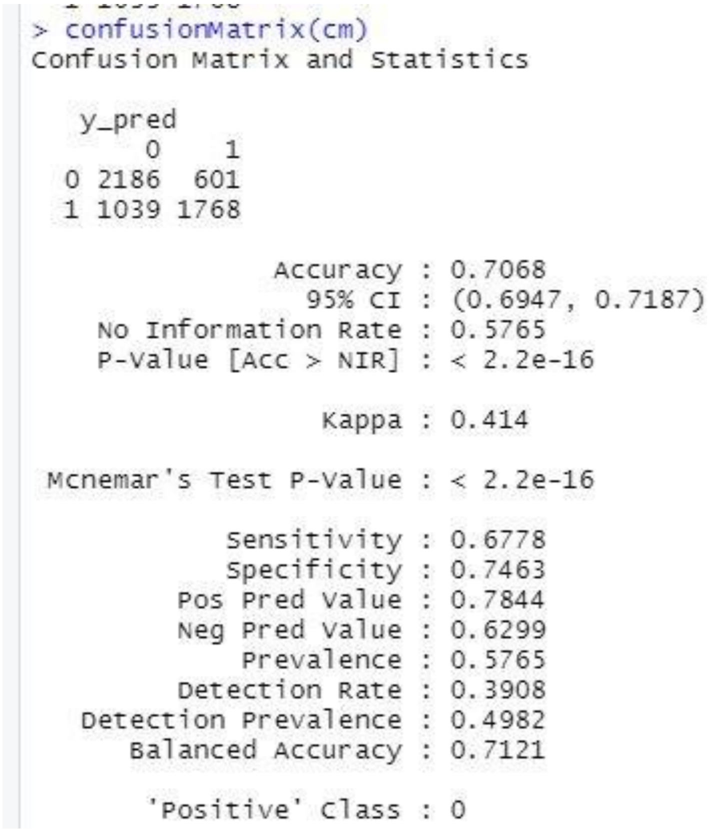
3.4 Support Vector Machine Model:

- The aim of the svm model is to take groups of observations and construct boundaries to predict which group feature observations belong to based on their measurements.
- The different groups that must be separated will be called classes.
- Svms can handle any number of classes, observations of any dimensions and svms can take any dimensions.
- For svm modelling, we are using the svm function from library e1071 to train the model from the trained dataset.
- Then we predict the data with a trained data module and to evaluate the algorithm we are using a confusion matrix.

3.4.1 Support Vector Machine Result.

The Support Vector Machine model achieves an accuracy up to 70.68%.

3.4.2 Support Vector Machine Evaluation.



```
> confusionMatrix(cm)
Confusion Matrix and Statistics

   y_pred
       0    1
  0 2186  601
  1 1039 1768

               Accuracy : 0.7068
                 95% CI : (0.6947, 0.7187)
    No Information Rate : 0.5765
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.414

 Mcnemar's Test P-Value : < 2.2e-16

            Sensitivity : 0.6778
            Specificity : 0.7463
         Pos Pred Value : 0.7844
         Neg Pred Value : 0.6299
             Prevalence : 0.5765
         Detection Rate : 0.3908
   Detection Prevalence : 0.4982
      Balanced Accuracy : 0.7121

       'Positive' Class : 0
```

Figure 22 confusion matrix for Support Vector Machine.

Precision of the support vector machine 76.52

Recall of the support vector machine 70.43

F1 score of the support vector machine 70.35

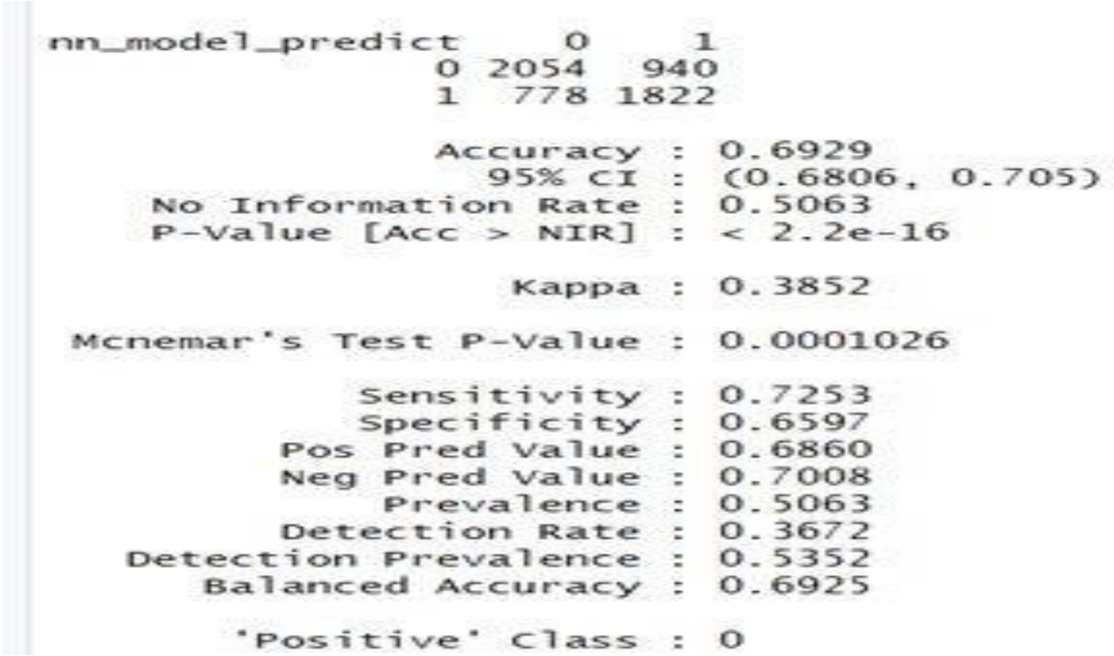
3.5 Neural Network Model:

- A **neural network** is a series of algorithms that endeavours to recognize underlying relationships in a set of data through a process that mimics the way the human brain operates.
- For neural network module, we are using the ‘nnet’ library package to train the dataset.
- Then we predict the trained dataset.
- For the performance analysis we are using the confusion matrix metric.

3.5.1 Neural Network Result:

The Neural Network model achieves an accuracy up to 69.29%.

3.5.2 Neural Network Evaluation:



```
nn_model_predict    0    1
               0 2054  940
               1  778 1822

               Accuracy : 0.6929
                 95% CI : (0.6806, 0.705)
    No Information Rate : 0.5063
    P-Value [Acc > NIR] : < 2.2e-16

                  Kappa : 0.3852

 Mcnemar's Test P-Value : 0.0001026

            Sensitivity : 0.7253
            Specificity : 0.6597
         Pos Pred Value : 0.6860
         Neg Pred Value : 0.7008
             Prevalence : 0.5063
         Detection Rate : 0.3672
   Detection Prevalence : 0.5352
      Balanced Accuracy : 0.6925

       'Positive' Class : 0
```

Figure 23 confusion matrix for Neural network model..

Precision of the neural networks 67.97

Recall of the neural networks 7043

F1 score of the neural networks 69.18

3.6 Decision Tree Model

- Decisions trees hold and display information in the form of a hierarchical structure that is easy to read and used for decision making.
- They are created using two steps: induction and pruning.
- We are using the rpart library package for modelling purposes.
- The training dataset is used to predict the model and then it is further evaluated using the test dataset and accuracy is determined from the confusion matrix.

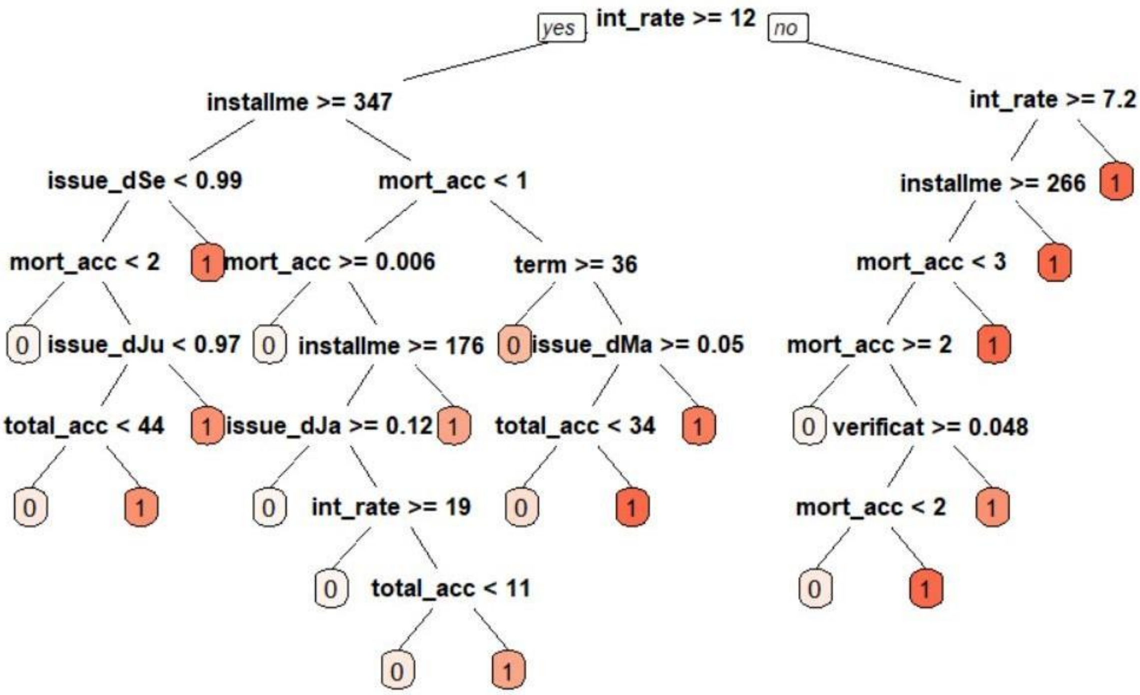
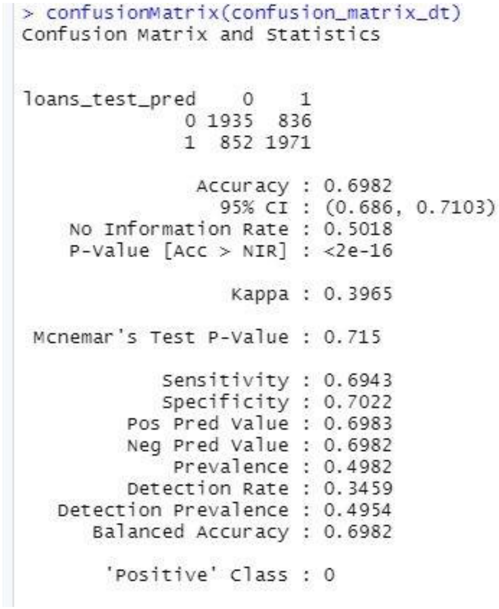


Figure 24 shows the graphical representation of the decision tree model.

3.6.1 Decision Tree Result

The decision tree model achieves an accuracy up to: 69.82

3.6.2 Decision Tree Evaluation:



```
> confusionMatrix(confusion_matrix_dt)
Confusion Matrix and Statistics


loans_test_pred    0    1
              0 1935  836
              1  852 1971

               Accuracy : 0.6982
                 95% CI : (0.686, 0.7103)
    No Information Rate : 0.5018
    P-Value [Acc > NIR] : <2e-16

                  Kappa : 0.3965

 Mcnemar's Test P-Value : 0.715

            Sensitivity : 0.6943
            Specificity : 0.7022
         Pos Pred Value : 0.6983
         Neg Pred Value : 0.6982
             Prevalence : 0.4982
         Detection Rate : 0.3459
   Detection Prevalence : 0.4954
      Balanced Accuracy : 0.6982

       'Positive' Class : 0
```

Figure 25 confusion matrix for Random Forest.

Precision of the decision tree 69.83

Recall of the decision tree 69.82

F1 score of the decision tree 69.83

3.6 K Nearest Neighbor Model:

K Nearest Neighbor algorithm classifies the data point on how its neighbor is classified. For calculating the nearest neighbor we use euclidean distance.

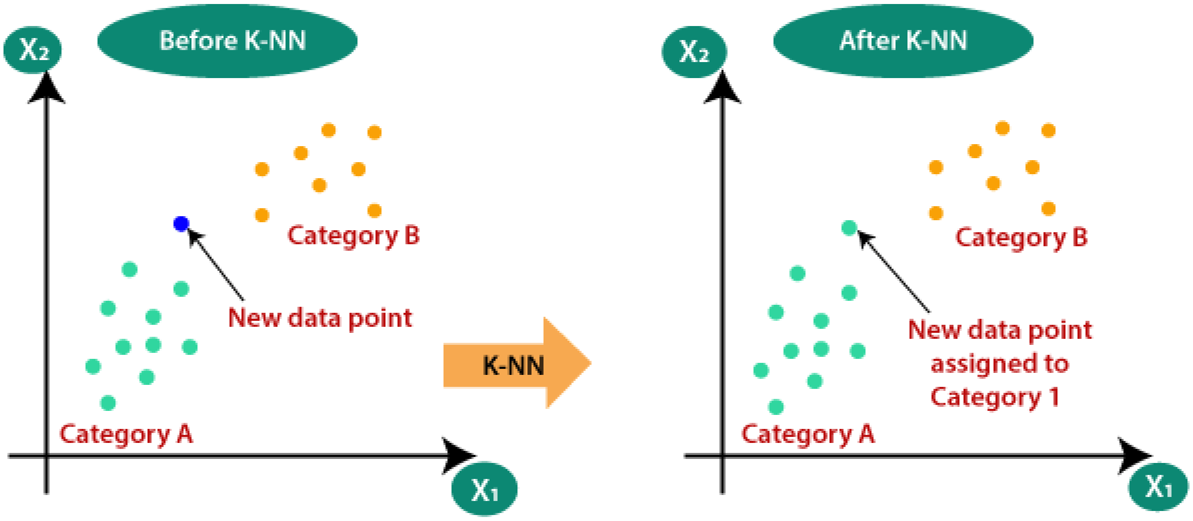


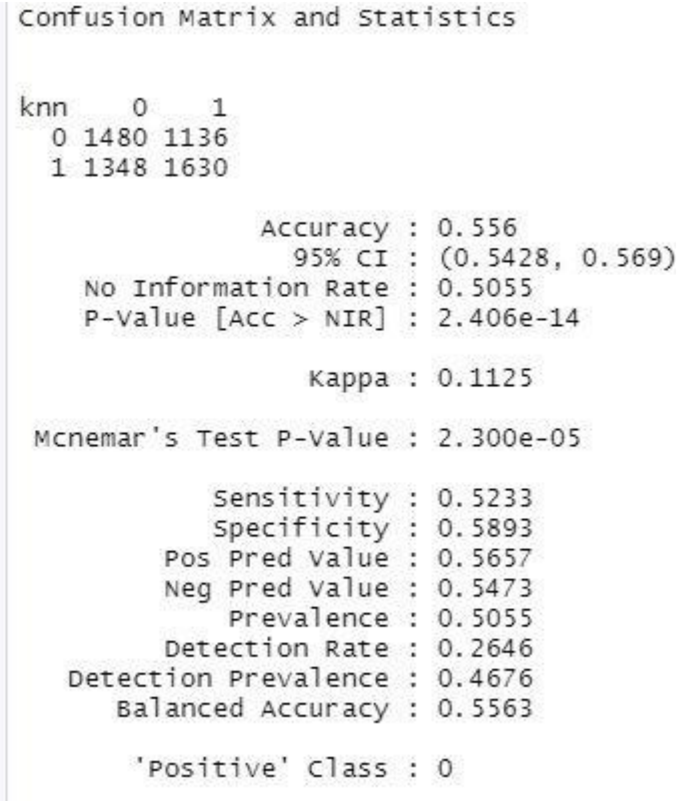
Figure 26, K in KNN represents the number of nearest neighbors we used to classify new data points. The ‘k’ value we selected was 114 was chosen based on the calculation of the square root of the total number of rows of data used to train the model.

3.3.1 K Nearest Neighbor Result:

The K nearest neighbor model achieves an accuracy up to: 55.6%

3.3.2 K Nearest Neighbor Evaluation:

Confusion Matrix



```
Confusion Matrix and Statistics


knn    0    1
  0 1480 1136
  1 1348 1630

               Accuracy : 0.556
                 95% CI : (0.5428, 0.569)
    No Information Rate : 0.5055
    P-Value [Acc > NIR] : 2.406e-14

                  Kappa : 0.1125

 Mcnemar's Test P-Value : 2.300e-05

            Sensitivity : 0.5233
            Specificity : 0.5893
         Pos Pred Value : 0.5657
         Neg Pred Value : 0.5473
             Prevalence : 0.5055
         Detection Rate : 0.2646
   Detection Prevalence : 0.4676
      Balanced Accuracy : 0.5563

       'Positive' Class : 0
```

Figure 27 confusion matrix for Neural network model..

Precision of the k nearest neighbor 56.5

Recall of the k nearest neighbor 52.3

F1 score of the k nearest neighbor 54.3

Section 4: Result and Interpretation

4.1 Table 3 represent summary of the result of the six models:

Model Name	Accuracy	Precision	Recall	F1 score
Random Forest	80.34%	80.41	81.89	81.15
Logistic Regression	68.82%	67.75	70.82	69.25
Support Vector Machine	70.68%	76.52	70.43	73.35
Neural Network	69.29%	67.97	70.43	69.18
K Nearest Neighbour	55.6%	56.5	52.3	54.3
Decision Trees	69.8%	69.83	69.82	69.83

Table 3 shows the summary of results of the six models.

4.2 Conclusion:

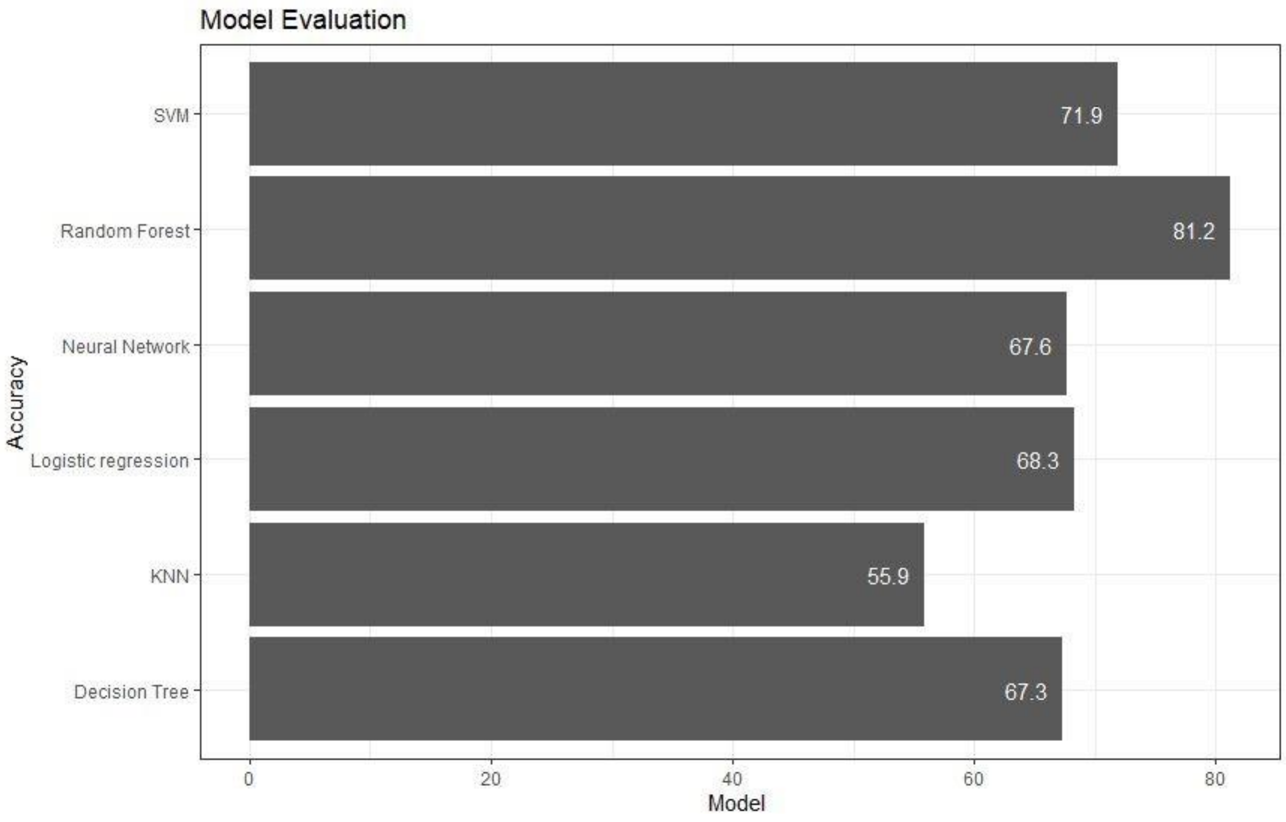


Figure 28, shows models evaluation

Out of all the models above, the random forest model delivers the best results. In our project, we need to correctly identify the borrowers who are not able to pay back the fully paid loan. When we go back to random forest results, we can see it is better in predicting the “Charged-off” value. In Random Forest, misclassification of “Charged-off” target variables is less than compared to other models. It can be shown in the evidence that in a random forest confusion matrix, we can clearly say that a positive predictive rate determines charged off variables. In our model “Charged-off “value is assigned as 0 and “Fully Paid” value is assigned as 1.

Assume “Charged off” as CO and “Fully Paid” as “FP”.

True Positive - Number of Correctly Predicting the CO as CO by the model
False Positive - Number of misclassifying the CO as FP by the model
Positive Predictive Rate = True Positive/(True Positive + False Positive)

Positive Predictive Rate (R.F model) = 2273/ (2273+514) = 0.8156

We can also see that the random forest model also correctly predicts the negative predictive value (correctly classifying the fully paid loan borrowers) with negative predictive value nearing 0.79. Another equivalent model which comes nearer to Random Forest is support vector machine which shows true positive rate as 0.79 but it is poor in determining negative predictive values (correctly classifying the fully paid loan borrowers).

Positive Predictive rate is also identified as precision and negative predictive rate is identified as recall.

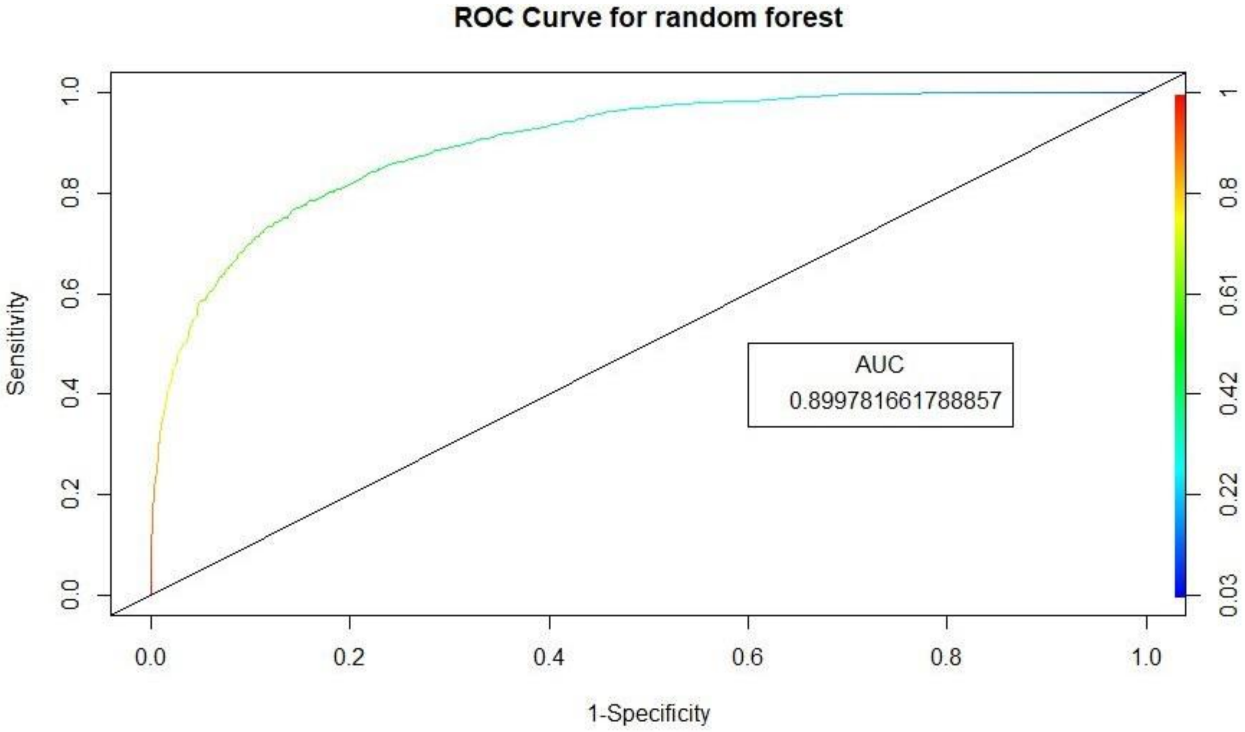


Figure 29 shows ROC Curve for random forest

ROC curve is a graphical representation of correctly identifying the borrower who got charged off and borrowers who fully paid the loan.The best ROC curve is the one which covers most area under the curve (AUC).After we plotter the ROC for random forest ,we can clearly see the random forest covers almost 89 percent of the curve.

To summarize, random forest is the best model amongst other models to identify whether the specific borrower is able to pay back the loan. Random forest gives best results not only in accuracy scores but also in other measurements like precision, recall and F1 score.

# The End #

# Thank you for reading #

# Thank you everyone for the efforts #