THE UNIVERSITY OF TEXAS AT DALLAS

PREDICTING LOAN DEFAULTERS USING DATA MINING METHOD

Master of Science in Business Analytics
Business Analytics with R, BUAN 6356, Spring 2022

Joyal Joby Chully

Sailesh Dogiparthi

Shubham.

Sriram Koushik Kanuri

Teffy Annie George

INDEX

ABSTRACT

In the lending industry, investors provide loans to borrowers in exchange for the promise of repayment with interest. If the borrower repays the loan, then the lender would make a profit from the interest. However, if the borrower fails to repay the loan, then the lender loses money. Therefore, lenders face the problem of predicting the risk of a borrower being unable to repay a loan. In this study, the data from the Lending club is used to train data mining models to determine if the borrower has the ability to repay their loan. As part of this, we would analyze the performance of the models (Decision Tree and Logistic Regression). As a result, the logistic regression model is found as the optimal predictive model and it is expected that Fico Score and annual income significantly influence the outcome.

Chapter 1: Introduction

1.1 Background

Peer-to-peer (P2P) lending enables individuals to obtain loans directly from other individuals, cutting out the financial institution as the middleman. Websites that facilitate P2P lending have greatly increased its adoption as an alternative method of financing.

P2P lending is also known as "social lending" or "crowdlending." It has only existed since 2005, but the crowd of competitors already includes Prosper, Lending Club, Upstart, and StreetShares.

P2P lending websites connect borrowers directly to investors. The site sets the rates and terms and enables the transactions. They are individual investors who want to get a better return on their cash savings than a bank savings account or Cash Deposit offers. The borrowers seek an alternative to traditional banks or a better rate than what banks offer.

Loan repayment is a common problem for lending companies. Whenever a customer asks for a loan at the bank, the bank faces the problem of failure of loan repayment. We will be using the dataset of Lending Club for our analysis. LendingClub is a peer-to-peer lender that offers personal loans through an online marketplace. The company assesses applicants' risk and lets investors lend directly to individuals or spread their money across a number of loans. It charges borrowers an origination fee of 1%-5% (depending on credit risk) and creditors a service fee equal to 1% of the loan amount. The company raised $172 million in 2013 from investors including Google Ventures, DST and Coatue Management.

1.2 Problem Statement

We know that the investors provide loans to the borrowers with the understanding that they will repay the loans with interest. Investors come across 2 sides. One where the borrower repays the loan. In this case, the lender has profited from the interest amount. On the other side, if a borrower is unable to pay the loan, the lender will lose this money. The lenders must face the problem of predicting this risk if a borrower is unable to repay the loan back.

We will analyze historical data and predict defaulters using Decision tree and Logistic regression models. We will also check the performance of these models using evaluators and decide if any model has dominance over the other.

Chapter 2: Dataset

2.1 Data Source

This dataset has 9578 records comprising the details of loans provided by LendingClub.com between the years 2007 and 2010. This data, taken from Kaggle will be used as we can now see that the users have already repaid or defaulted. We will apply data mining methods to loan default predictions which will be helpful to solve business problems.

2.2 Features

not.fully.paid: This is the dependent variable which indicates if the loan was paid back to the lender by the borrower in full or not (borrower can either be marked as default or the borrower was unlikely to pay it back)

Below are the marked independent variables:

credit.policy: 1 if the customer meets the credit underwriting criteria of LendingClub.com, and 0 otherwise. This is based on an evaluation done by the lending entity as per lending protocols

purpose: the purpose of loan (can be 'credit_card', 'educational', 'major_purchase','small_business','debt_consolidation','all_other', 'home improvement')

int.rate: The interest rate of the loan, as a proportion (a rate of 11% would be stored as 0.11). Borrowers judged by LendingClub.com to be riskier are assigned higher interest rates.

installment: The monthly installments ($) owed by the borrower if the loan is funded.

log.annual.inc: The natural log of the self-reported annual income of the borrower.

dti: The debt-to-income ratio of the borrower (amount of debt divided by annual income).

fico: The FICO credit score of the borrower. A FICO Score is a three-digit number based on the information in one's credit reports

days.with.cr.line: The number of days the borrower has had a credit line.

revol.bal: The borrower's revolving balance (amount unpaid at the end of the credit card billing cycle).

revol.util: The borrower's revolving line utilization rate (the amount of the credit line used relative to total credit available).

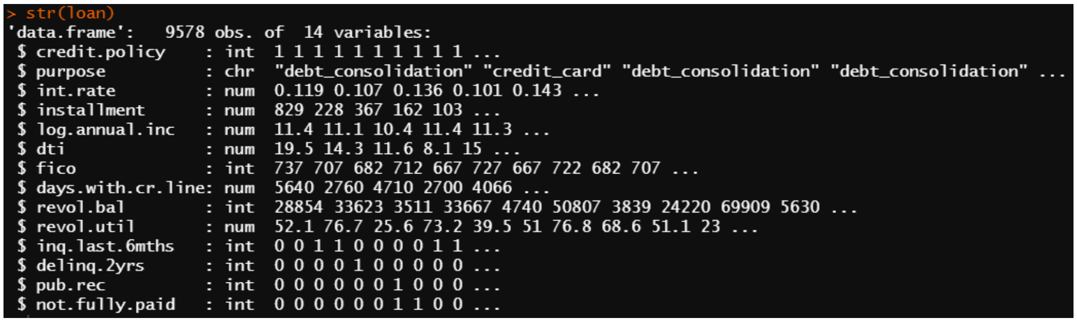
inq.last.6mths: The borrower's number of inquiries by creditors in the last 6 months.

delinq.2yrs: The number of times the borrower had been 30+ days past due on a payment in the past 2 years.

pub.rec: The borrower's number of derogatory public records (bankruptcy filings, tax liens, or judgments).

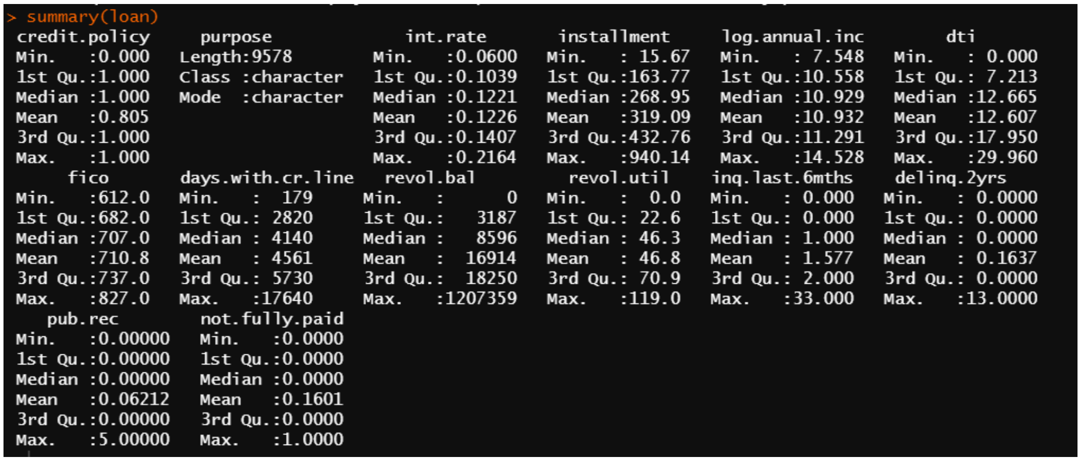
2.3 Data Exploration (EDA)

Exploratory Data Analysis is an approach to data analysis that postpones the usual assumptions about what kind of model the data follow with the more direct approach of allowing the data itself to reveal its underlying structure and model. We checked the structure of the data set to check the data types of variables.



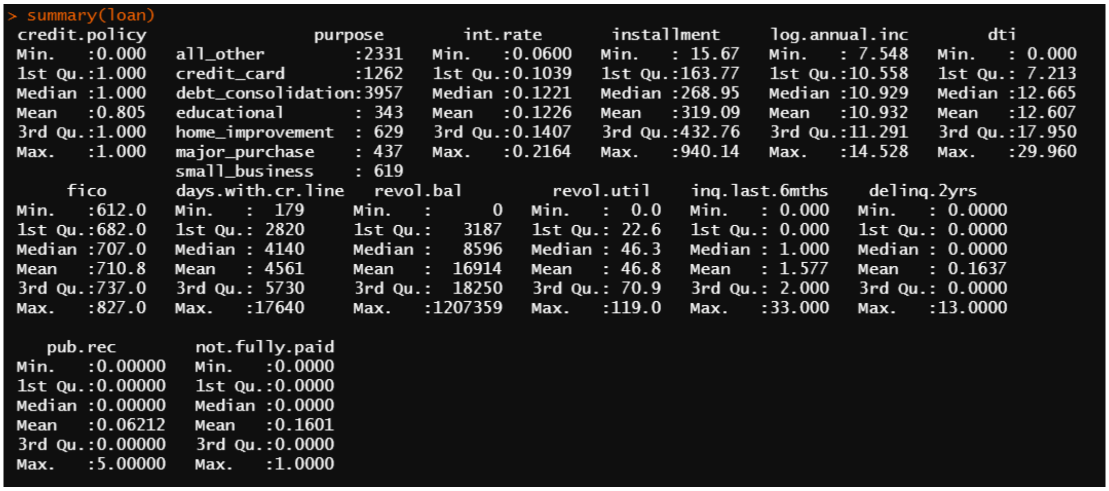
```
> str(loan)
'data.frame':   9578 obs. of  14 variables:
 $ credit.policy    : int  1 1 1 1 1 1 1 1 1 1 ...
 $ purpose          : chr  "debt_consolidation" "credit_card" "debt_consolidation" "debt_consolidation" ...
 $ int.rate         : num  0.119 0.107 0.136 0.101 0.143 ...
 $ installment      : num  829 228 367 162 103 ...
 $ log.annual.inc   : num  11.4 11.1 10.4 11.4 11.3 ...
 $ dti              : num  19.5 14.3 11.6 8.1 15 ...
 $ fico             : int  737 707 682 712 667 727 667 722 682 707 ...
 $ days.with.cr.line: num  5640 2760 4710 2700 4066 ...
 $ revol.bal        : int  28854 33623 3511 33667 4740 50807 3839 24220 69909 5630 ...
 $ revol.util       : num  52.1 76.7 25.6 73.2 39.5 51 76.8 68.6 51.1 23 ...
 $ inq.last.6mths   : int  0 0 1 1 0 0 0 0 1 1 ...
 $ delinq.2yrs      : int  0 0 0 0 1 0 0 0 0 0 ...
 $ pub.rec          : int  0 0 0 0 0 0 1 0 0 0 ...
 $ not.fully.paid   : int  0 0 0 0 0 0 1 1 0 0 ...
```

As we can observe, 7 variables in int, 6 variables in num, 1 variable in char



```
> summary(loan)
 credit.policy     purpose             int.rate       installment      log.annual.inc        dti        
 Min.   :0.000   Length:9578        Min.   :0.0600   Min.   : 15.67   Min.   : 7.548   Min.   : 0.000  
 1st Qu.:1.000   Class :character   1st Qu.:0.1039   1st Qu.:163.77   1st Qu.:10.558   1st Qu.: 7.213  
 Median :1.000   Mode  :character   Median :0.1221   Median :268.95   Median :10.929   Median :12.665  
 Mean   :0.805                      Mean   :0.1226   Mean   :319.09   Mean   :10.932   Mean   :12.607  
 3rd Qu.:1.000                      3rd Qu.:0.1407   3rd Qu.:432.76   3rd Qu.:11.291   3rd Qu.:17.950  
 Max.   :1.000                      Max.   :0.2164   Max.   :940.14   Max.   :14.528   Max.   :29.960  
      fico       days.with.cr.line   revol.bal         revol.util    inq.last.6mths    delinq.2yrs    
 Min.   :612.0   Min.   :  179     Min.   :      0   Min.   :  0.0   Min.   : 0.000   Min.   : 0.0000  
 1st Qu.:682.0   1st Qu.: 2820     1st Qu.:   3187   1st Qu.: 22.6   1st Qu.: 0.000   1st Qu.: 0.0000  
 Median :707.0   Median : 4140     Median :   8596   Median : 46.3   Median : 1.000   Median : 0.0000  
 Mean   :710.8   Mean   : 4561     Mean   :  16914   Mean   : 46.8   Mean   : 1.577   Mean   : 0.1637  
 3rd Qu.:737.0   3rd Qu.: 5730     3rd Qu.:  18250   3rd Qu.: 70.9   3rd Qu.: 2.000   3rd Qu.: 0.0000  
 Max.   :827.0   Max.   :17640     Max.   :1207359   Max.   :119.0   Max.   :33.000   Max.   :13.0000  
    pub.rec        not.fully.paid  
 Min.   :0.00000   Min.   :0.0000  
 1st Qu.:0.00000   1st Qu.:0.0000  
 Median :0.00000   Median :0.0000  
 Mean   :0.06212   Mean   :0.1601  
 3rd Qu.:0.00000   3rd Qu.:0.0000  
 Max.   :5.00000   Max.   :1.0000  
```

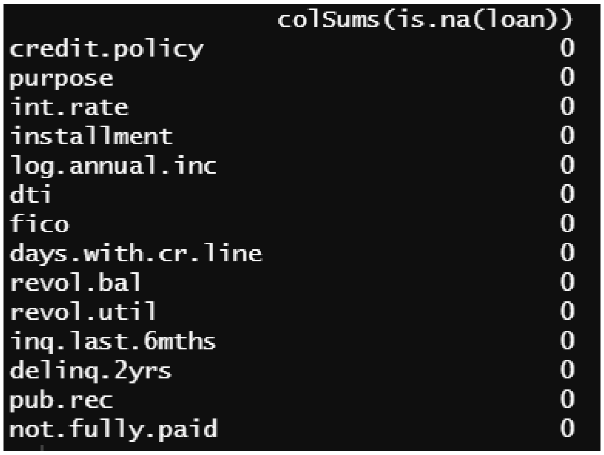
As we can observe, “purpose” variable is a categorical variable and hence we are not getting the summary, to get some idea about this variable’s summary, we can convert it to factor class.



```
> summary(loan)
 credit.policy                 purpose        int.rate        installment      log.annual.inc        dti        
 Min.   :0.000   all_other         :2331   Min.   :0.0600   Min.   : 15.67   Min.   : 7.548   Min.   : 0.000  
 1st Qu.:1.000   credit_card       :1262   1st Qu.:0.1039   1st Qu.:163.77   1st Qu.:10.558   1st Qu.: 7.213  
 Median :1.000   debt_consolidation:3957   Median :0.1221   Median :268.95   Median :10.929   Median :12.665  
 Mean   :0.805   educational       : 343   Mean   :0.1226   Mean   :319.09   Mean   :10.932   Mean   :12.607  
 3rd Qu.:1.000   home_improvement  : 629   3rd Qu.:0.1407   3rd Qu.:432.76   3rd Qu.:11.291   3rd Qu.:17.950  
 Max.   :1.000   major_purchase    : 437   Max.   :0.2164   Max.   :940.14   Max.   :14.528   Max.   :29.960  
                 small_business    : 619                                                                    
      fico      days.with.cr.line   revol.bal         revol.util    inq.last.6mths    delinq.2yrs    
 Min.   :612.0   Min.   :  179    Min.   :      0   Min.   :  0.0   Min.   : 0.000   Min.   : 0.0000  
 1st Qu.:682.0   1st Qu.: 2820    1st Qu.:   3187   1st Qu.: 22.6   1st Qu.: 0.000   1st Qu.: 0.0000  
 Median :707.0   Median : 4140    Median :   8596   Median : 46.3   Median : 1.000   Median : 0.0000  
 Mean   :710.8   Mean   : 4561    Mean   :  16914   Mean   : 46.8   Mean   : 1.577   Mean   : 0.1637  
 3rd Qu.:737.0   3rd Qu.: 5730    3rd Qu.:  18250   3rd Qu.: 70.9   3rd Qu.: 2.000   3rd Qu.: 0.0000  
 Max.   :827.0   Max.   :17640    Max.   :1207359   Max.   :119.0   Max.   :33.000   Max.   :13.0000  

    pub.rec        not.fully.paid  
 Min.   :0.00000   Min.   :0.0000  
 1st Qu.:0.00000   1st Qu.:0.0000  
 Median :0.00000   Median :0.0000  
 Mean   :0.06212   Mean   :0.1601  
 3rd Qu.:0.00000   3rd Qu.:0.0000  
 Max.   :5.00000   Max.   :1.0000  
```

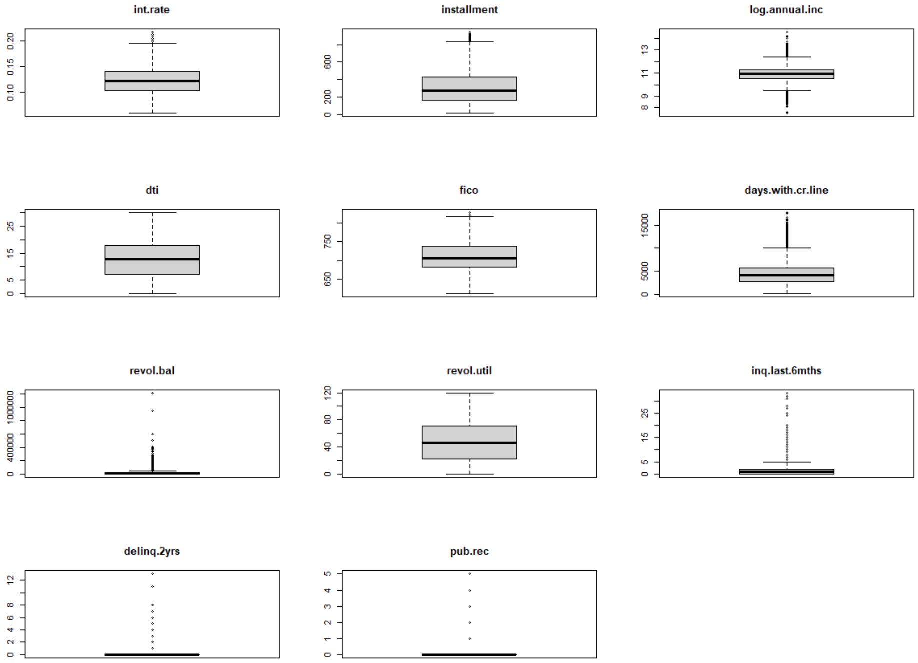
We checked if the data set has any missing values. There are 2 approaches to deal missing values. We can either drop the values with null in the dataset or impute the value with their means/medians.



```
                   colSums(is.na(loan))
credit.policy                         0
purpose                               0
int.rate                              0
installment                           0
log.annual.inc                        0
dti                                   0
fico                                  0
days.with.cr.line                     0
revol.bal                             0
revol.util                            0
inq.last.6mths                        0
delinq.2yrs                           0
pub.rec                               0
not.fully.paid                        0
```

There are no missing values found in the dataset.

We can check for outliers with a boxplot. As we can observe below, we are having some outliers in a few variables.



2.3.1 Correlation

Based on Fig 2.3.1, only a few of the variables seem to be correlated with others. For example, the feature “interest.rate” and “revol.util” are highly correlated, and the feature “credit.policy” and the “fico” are somehow correlated. ‘fico’ and ‘int.rate’ show a strong negative correlation. We will therefore,use these field in our final analysis to predict our outcome.

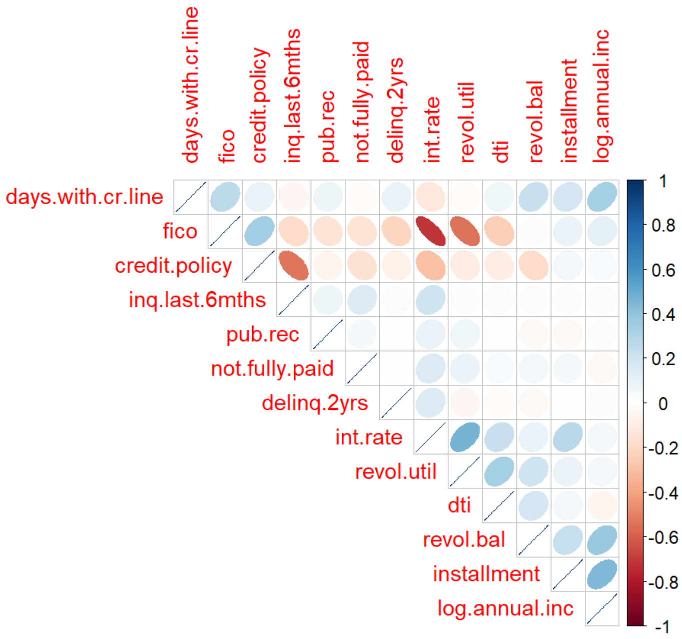


Figure 2.3.1: Correlation Plot

2.3.2 Imbalanced Data

Based on Fig 2.3.2, we can see that the data is imbalanced. The not.fully.paid variable is the dependent variable with which the independent variables need to be compared with. We can see 16% as those who have not paid the loan fully. Therefore, we cannot depend on accuracy alone in this case as when we predict, it will all belong to the negative class of 84% shown below.

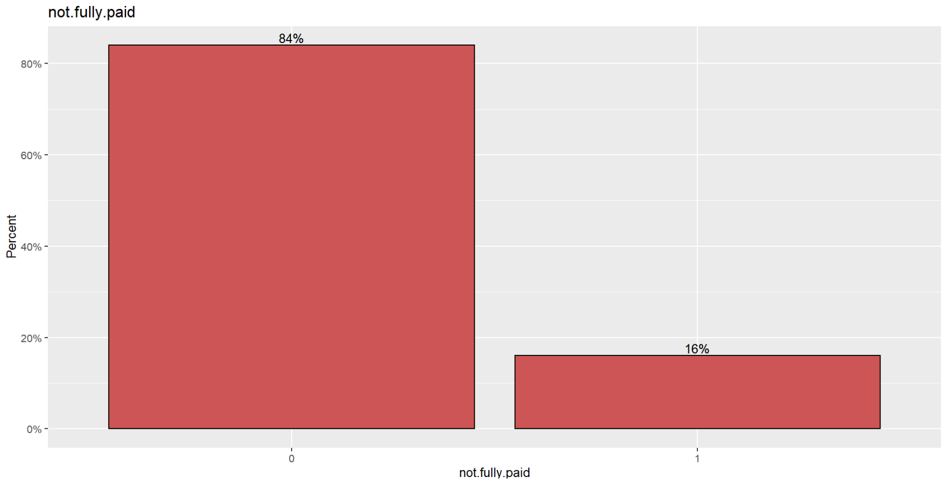


Figure 2.3.2: Frequency Distribution of dependent variable, not. fully.paid

2.3.3 Categorical Feature

The dataset has 2 categorical feature, “purpose” and “credit.policy” other than not.fully.paid. For the categorical feature ‘purpose’ , there are 7 possible categories: “credit card”, “major purchase”, “home improvement”, “educational”, “debt consolidation”, “small business” and “all others”.Therefore, the feature ”purpose” is set to factor since it is a nominal (not ordinal) categorical variable.

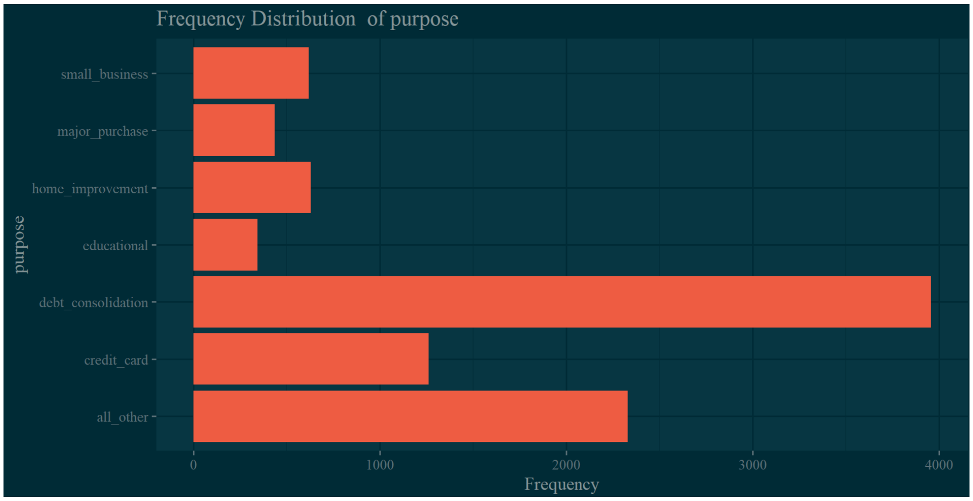


Figure 2.3.3: Frequency Distribution of categorical variable, purpose

2.3.4 Other Plots

The below graphs show the density and bar plots of independent variables. Based on the data size and data values, we will suggest data scaling so as to normalize the data.

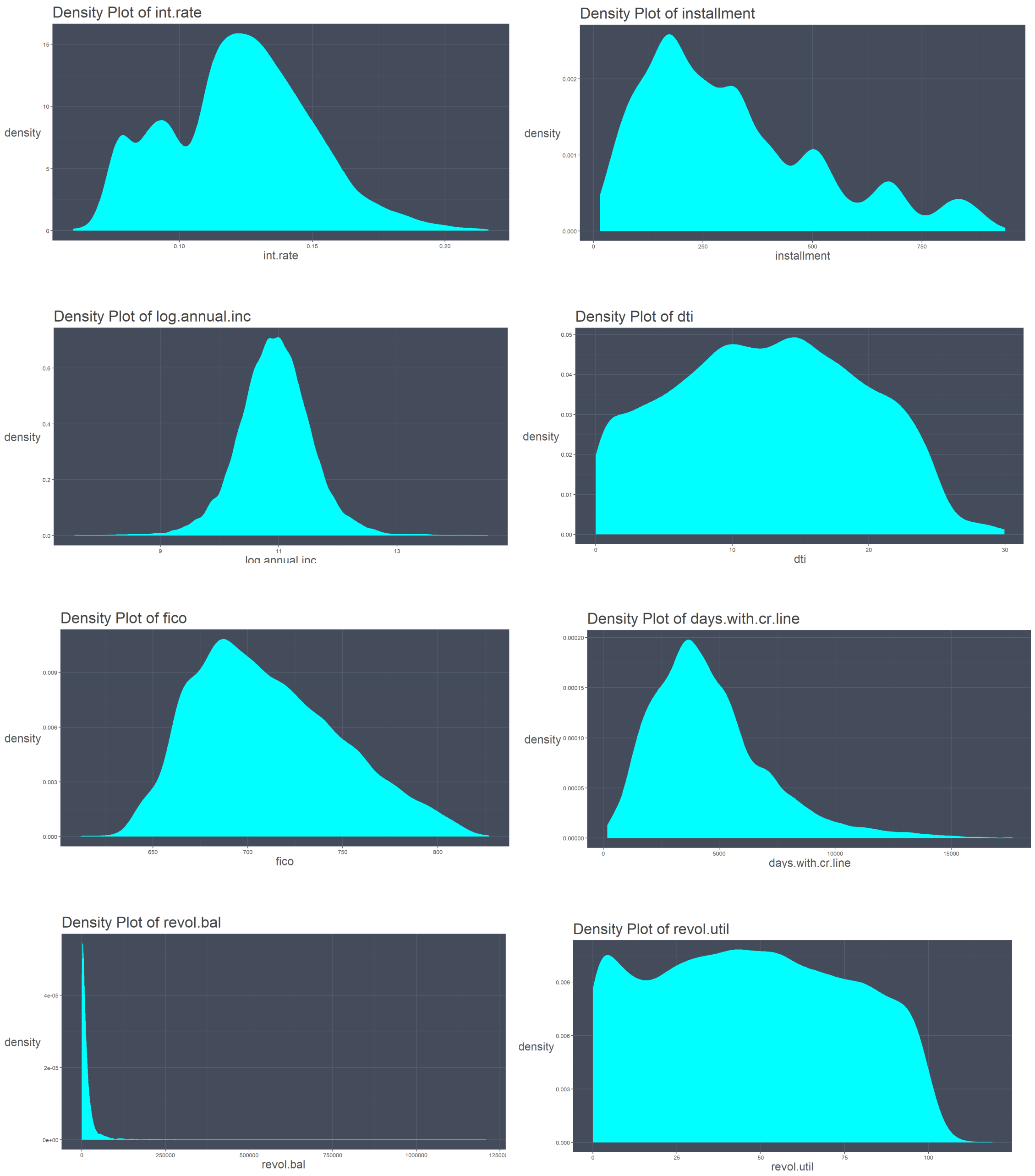


Figure 2.3.4: Density Plots of independent variables

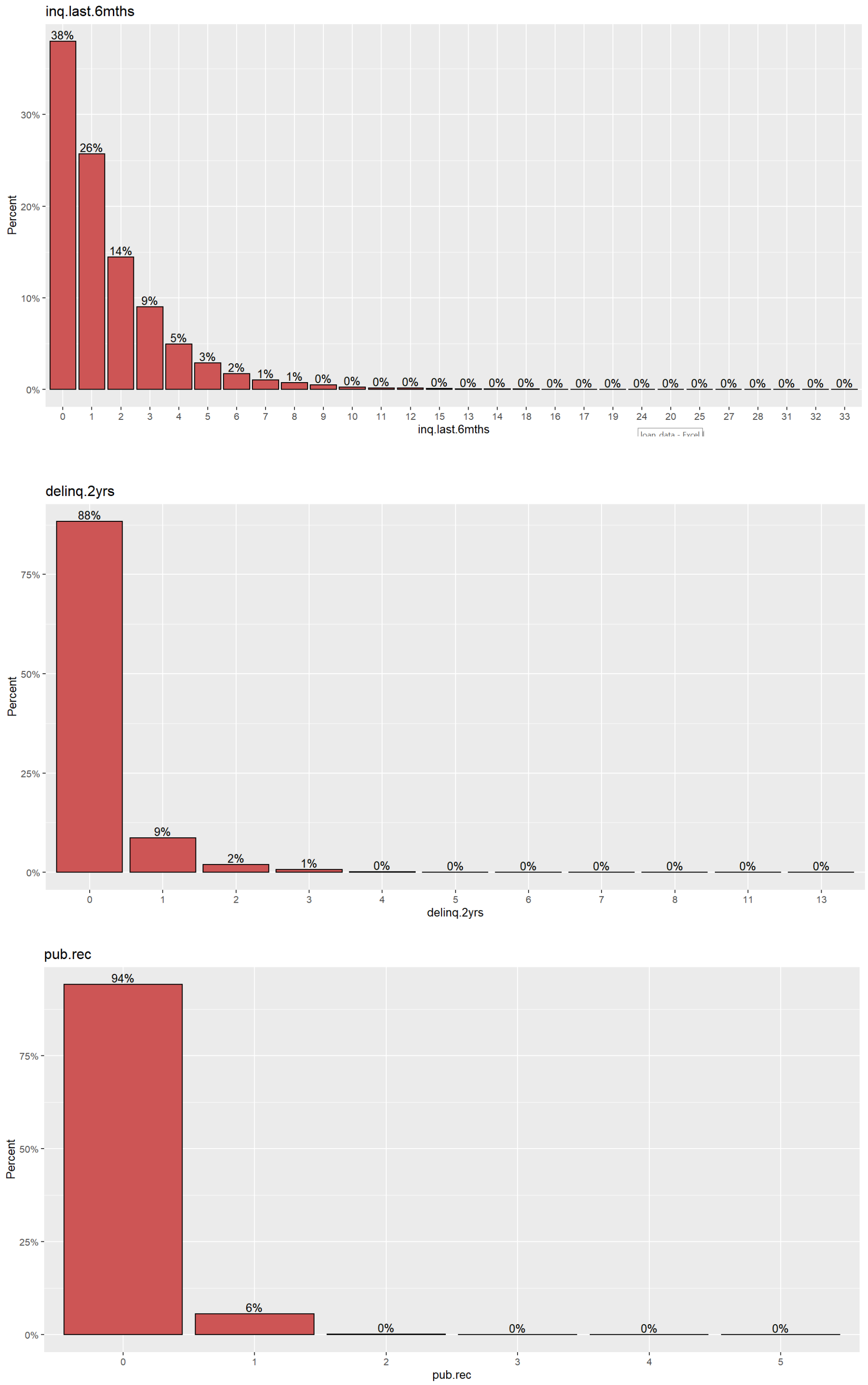


Figure 2.3.4: Distribution of independent variables

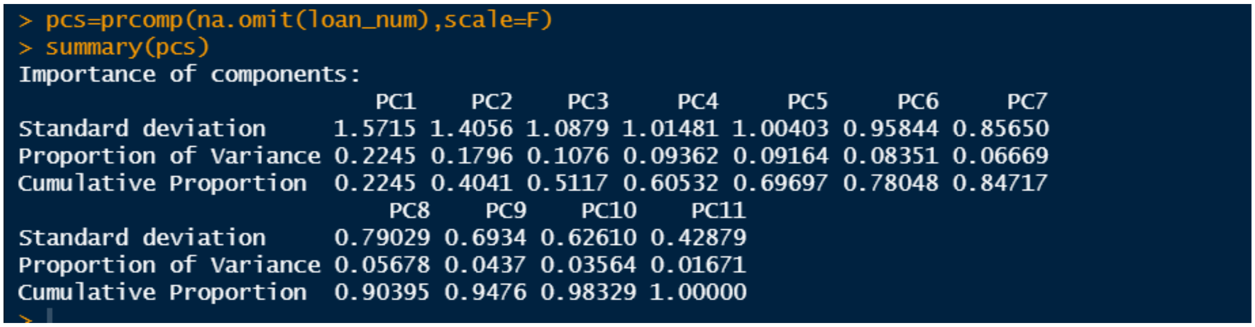
2.4 Data Standardization

Standardization of datasets is a common requirement for many data mining methods. It is possible that the model we created behaves badly if numerical data values have significant differences from others. Therefore, in order to improve the result of the prediction model, we will standardize the data first.

The main advantage of standardization scaling is to avoid attributes in greater numeric ranges dominating those in smaller numeric ranges. Another advantage is to avoid numerical difficulties during the calculation.

2.5 Dimension Reduction

Dimensionality reduction refers to techniques for reducing the number of input variables in training data. We have used Principal Component Analysis(PCA) to check the possibility of reducing the numerical variables by checking the principal components associated with it. Based on PCA, we try to reduce the dimensionality of the data and increase the interpretability without missing any information. However, in performing this reduction method, PCA suggests the usage of all numerical variables for modelling.



```
> pcs=prcomp(na.omit(loan_num),scale=F)
> summary(pcs)
Importance of components:
                          PC1    PC2    PC3     PC4     PC5     PC6     PC7
Standard deviation     1.5715 1.4056 1.0879 1.01481 1.00403 0.95844 0.85650
Proportion of Variance 0.2245 0.1796 0.1076 0.09362 0.09164 0.08351 0.06669
Cumulative Proportion  0.2245 0.4041 0.5117 0.60532 0.69697 0.78048 0.84717
                           PC8    PC9    PC10    PC11
Standard deviation     0.79029 0.6934 0.62610 0.42879
Proportion of Variance 0.05678 0.0437 0.03564 0.01671
Cumulative Proportion  0.90395 0.9476 0.98329 1.00000
>
```

Chapter 3: MODELLING

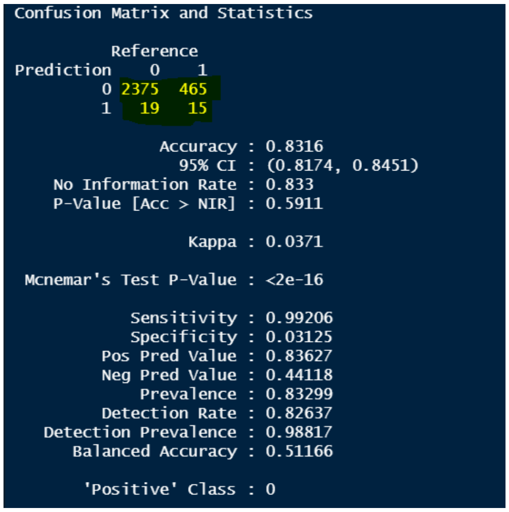
3.1 Introduction

We are preparing the classification model which is classified under supervised learning. Supervised learning is the task of learning a function that maps an input to an output based on example input-output pairs. We are checking here if the lender is able to repay the money back or not. We will analyze these using classification models: Decision Tree and Logistic Regression.

The classification model approach requires a good quality of the training set data. We will explore the loan repayment data set and perform data exploratory analysis and fine-tune the data for a better predictive outcome. The data is split into 2 sets: the training set (70%) and the validation set (30%). We will use the training set to fit the model and the validation set to evaluate the model performance.

Since we are dealing with imbalanced data, we will take into consideration the area under Receiver Operating Characteristics(ROC) curve, Precision and Recall instead of the accuracy score to measure the performance of the model. We are not evaluating our models based on accuracy because, even if the model predicts all individuals to be defaulters, we will get 84% accuracy (because of imbalance in data), even when there are cases otherwise in historical data. This sort of model is deemed to be counter-intuitive. Class imbalance methods like the ROC curve give a more rule biased towards the majority class.

Here, we can see that for the not.fully.paid field, the data is more on the true negative side(2375) as to which accuracy will be high. As to this imbalance in data, we will consider class-imbalance methods to determine a good classifier.



```
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2375  465
         1   19   15

               Accuracy : 0.8316
                 95% CI : (0.8174, 0.8451)
    No Information Rate : 0.833
    P-Value [Acc > NIR] : 0.5911

                  Kappa : 0.0371

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99206
            Specificity : 0.03125
         Pos Pred Value : 0.83627
         Neg Pred Value : 0.44118
             Prevalence : 0.83299
         Detection Rate : 0.82637
   Detection Prevalence : 0.98817
      Balanced Accuracy : 0.51166

       'Positive' Class : 0
```

Precision = True Positives/(True Positives + False Positives)

Recall = True Positives/(True Positives + False Negatives)

Precision = 2375/(2375 + 465) = 0.83627

Recall/Sensitivity = 2375/(2375 + 19) = 0.99206

3.2 Decision Tree

A decision tree is a special type of flowchart used to visualize the decision-making process by mapping out different courses of action as well as their possible outcomes. Growing a tree involves deciding on which features to choose and what conditions to use for splitting, along with knowing when to stop. The tree is drawn upside down with its root at the top.

We have modelled the decision tree with the loan data. We evaluated the model with the validation data and received the below test results. In the below-plotted Fig 3.2.2., we have set the number of observations to split as 10. The improvement set at each node is 0.002

Precision : 0.8350

Recall: 0.9937

AUC Curve: 0.5874

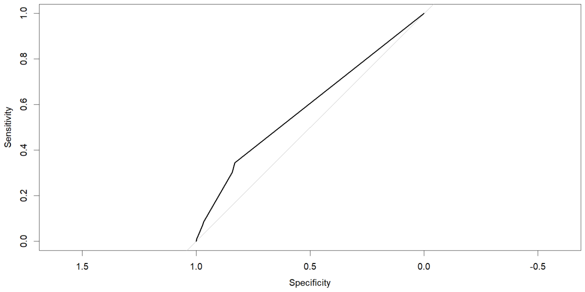
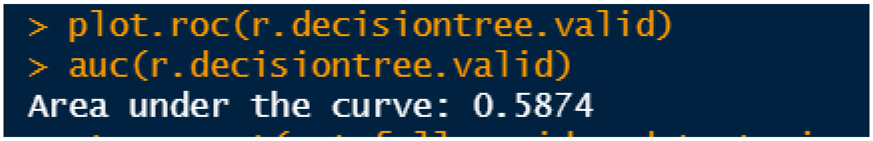
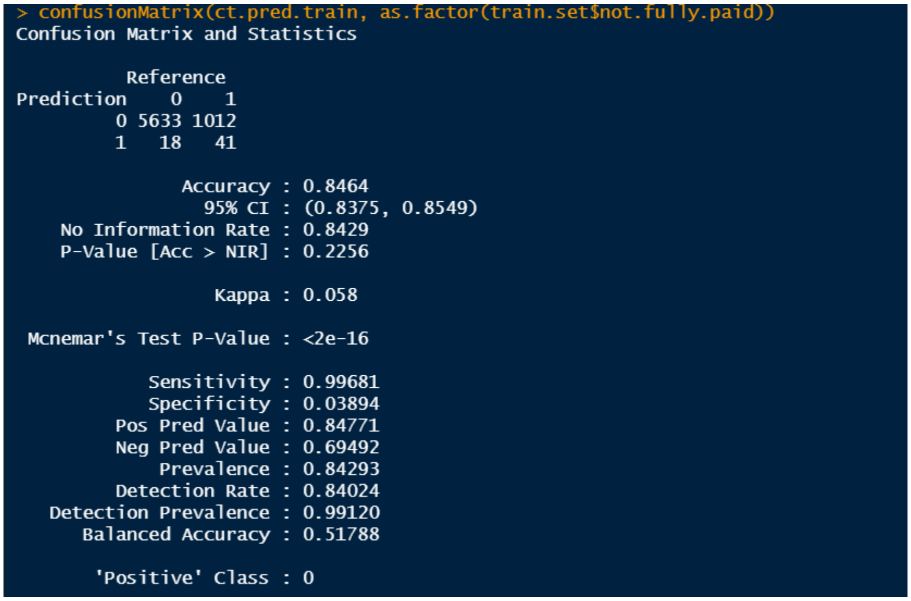


Figure 3.2.1: AUC Curve Chart for Decision Tree



```
> plot.roc(r.decisiontree.valid)
> auc(r.decisiontree.valid)
Area under the curve: 0.5874
```



```
> confusionMatrix(ct.pred.train, as.factor(train.set$not.fully.paid))
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 5633 1012
         1   18   41

               Accuracy : 0.8464
                 95% CI : (0.8375, 0.8549)
    No Information Rate : 0.8429
    P-Value [Acc > NIR] : 0.2256

                  Kappa : 0.058

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99681
            Specificity : 0.03894
         Pos Pred Value : 0.84771
         Neg Pred Value : 0.69492
             Prevalence : 0.84293
         Detection Rate : 0.84024
   Detection Prevalence : 0.99120
      Balanced Accuracy : 0.51788

       'Positive' Class : 0
```

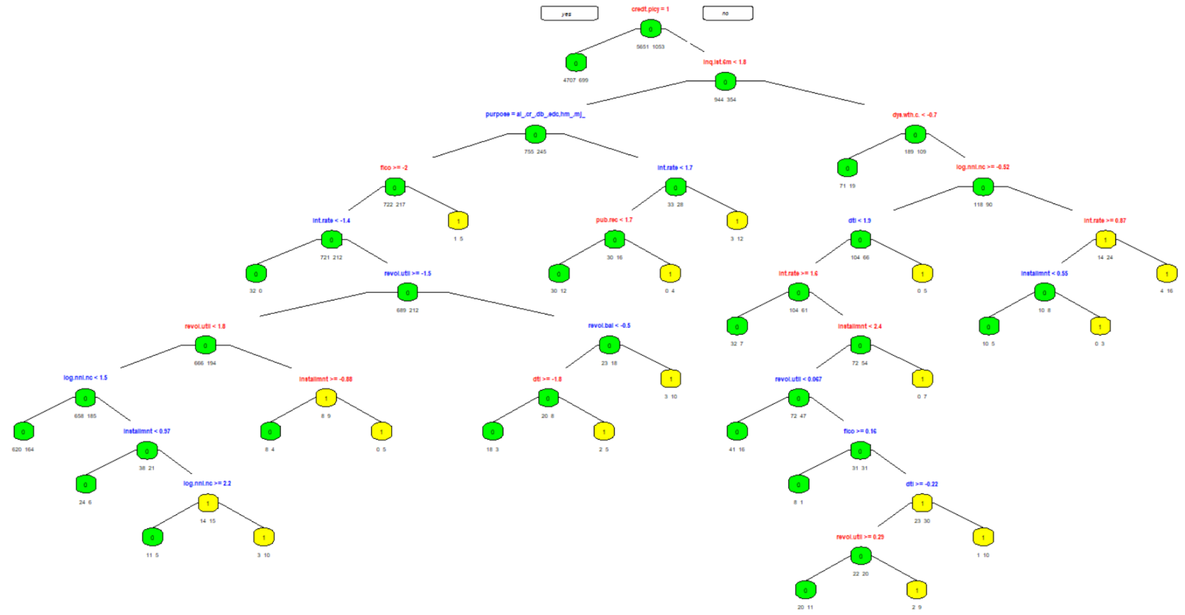
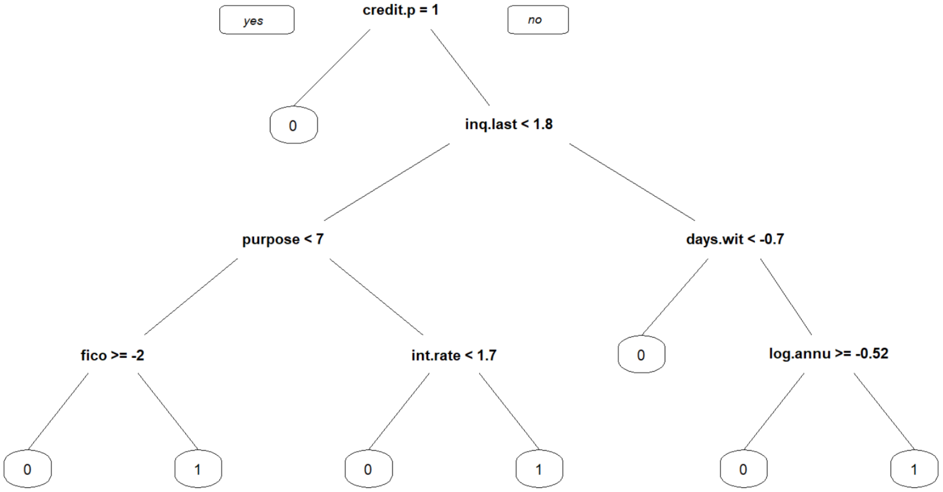


Figure 3.2.2: Decision Tree



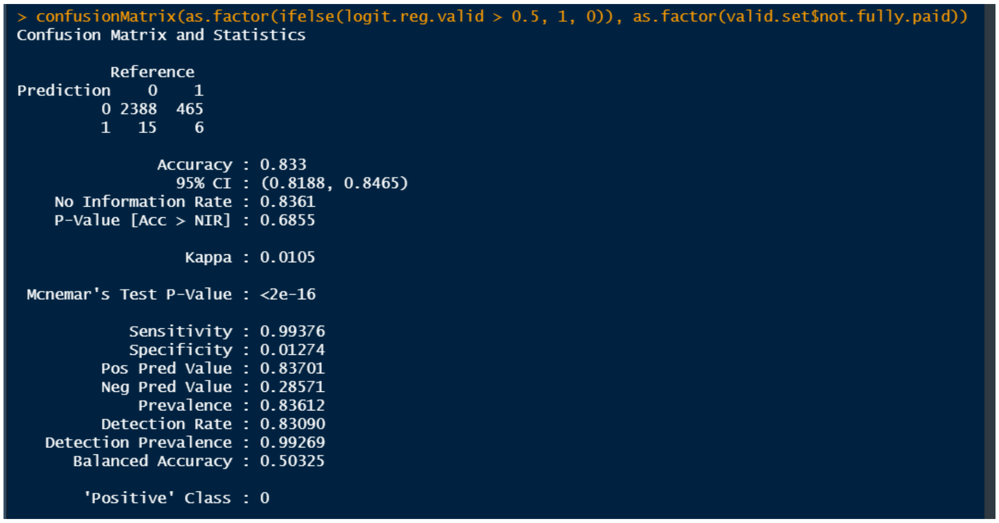
3.3 Logistic Regression

Logistic regression is a process of modelling the probability of a discrete outcome given an input variable. The most common logistic regression model gives a binary outcome; such as true/false, yes/no, etc. It is commonly used to obtain the odds ratio in the presence of more than one explanatory variable.

We will fit the logistic model into the loan data. We have created a model using the training data and performed a confusion matrix to check how the model performs.

We will check the accuracy using the confusion matrix and we check the ROC curve as this analysis is scale-invariant.

ROC curve can help in deciding the best threshold value. A ROC curve is plotted with sensitivity on the X-axis and (1 – specificity) on the y-axis. A high threshold value gives - high specificity and low sensitivity. A low threshold value gives - low specificity and high sensitivity. The closer the AUC is to 1, the better the model.



```
> confusionMatrix(as.factor(ifelse(logit.reg.valid > 0.5, 1, 0)), as.factor(valid.set$not.fully.paid))
Confusion Matrix and Statistics

          Reference
Prediction    0    1
         0 2388  465
         1   15    6

               Accuracy : 0.833
                 95% CI : (0.8188, 0.8465)
    No Information Rate : 0.8361
    P-Value [Acc > NIR] : 0.6855

                  Kappa : 0.0105

 Mcnemar's Test P-Value : <2e-16

            Sensitivity : 0.99376
            Specificity : 0.01274
         Pos Pred Value : 0.83701
         Neg Pred Value : 0.28571
             Prevalence : 0.83612
         Detection Rate : 0.83090
   Detection Prevalence : 0.99269
      Balanced Accuracy : 0.50325

       'Positive' Class : 0
```

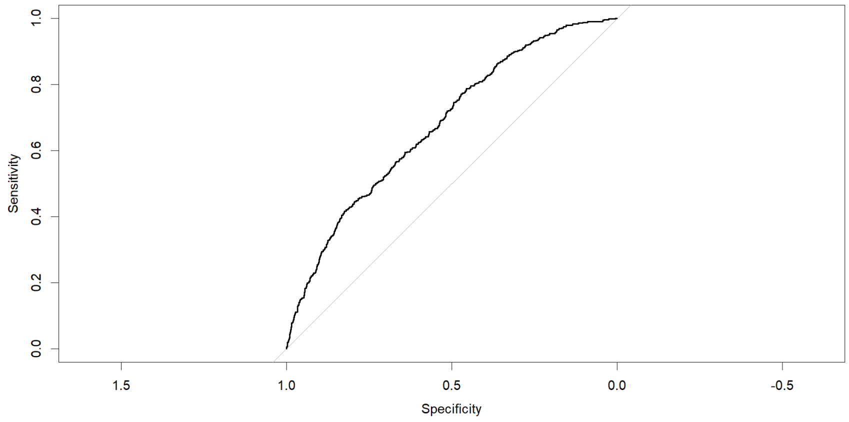
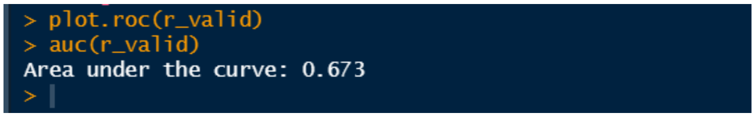


Figure 3.3.1: AUC Curve Chart for Logistic Tree

Precision : 0.8350

Recall: 0.9962

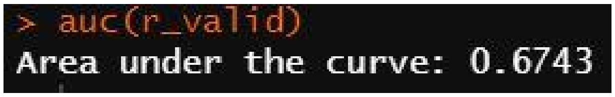
AUC Curve: 0.673



```
> plot.roc(r_valid)
> auc(r_valid)
Area under the curve: 0.673
>
```

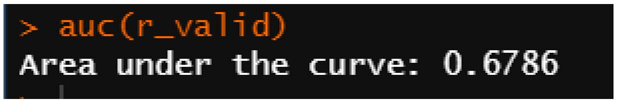
We tried eliminating numerical variables one by one and checked the AUC value for the said scenarios. Based on the below table, we could see that there was not much improvement even on eliminating as to which we are concluding to include all the variables (default case) as the valid model.

After removing int.rate from standardized data frame



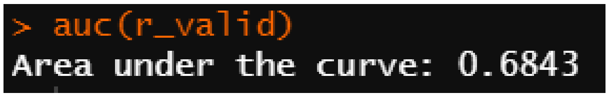
```
> auc(r_valid)
Area under the curve: 0.6743
```

After removing int.rate and dti from standardized data frame



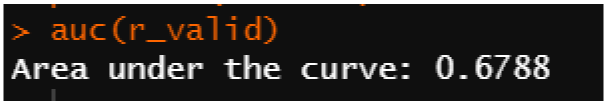
```
> auc(r_valid)
Area under the curve: 0.6786
```

After removing int.rate,dti and days.with.cr.line from standardized dataframe



```
> auc(r_valid)
Area under the curve: 0.6843
```

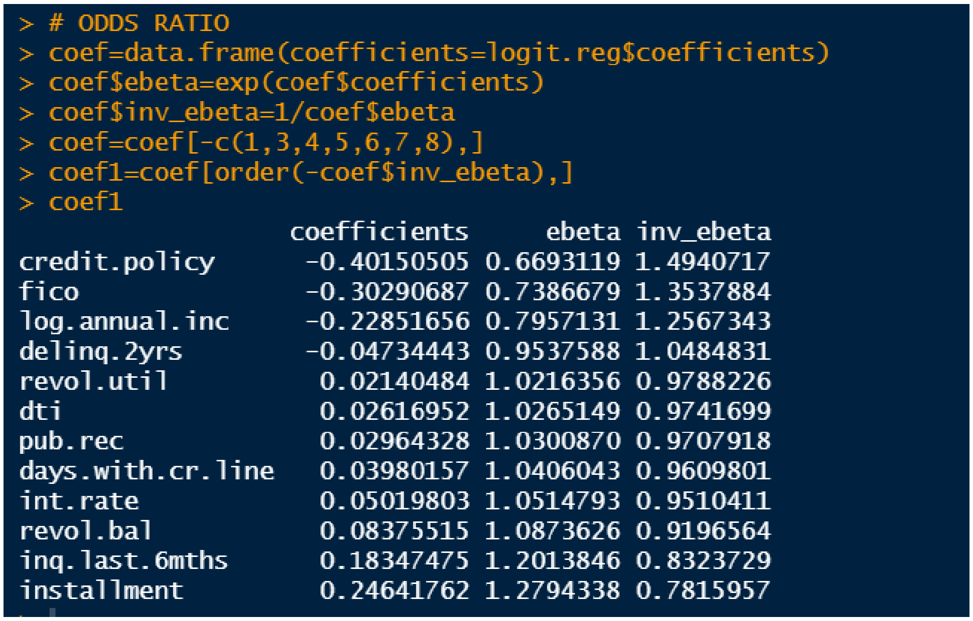
After removing int.rate, dti, days.with.cr.line and revol.util from standardized dataframe



```
> auc(r_valid)
Area under the curve: 0.6788
```

3.3.1 Odds Ratio

An odds ratio (OR) is a statistic that quantifies the strength of the association between two events. We have calculated the odds ratio for the logistic model to check if the loan borrower is likely to pay the loan back to the lending club. We can see that credit.policy, fico and annual income show field with high value and installment as the lowest value.



```
> # ODDS RATIO
> coef=data.frame(coefficients=logit.reg$coefficients)
> coef$ebeta=exp(coef$coefficients)
> coef$inv_ebeta=1/coef$ebeta
> coef=coef[-c(1,3,4,5,6,7,8),]
> coef1=coef[order(-coef$inv_ebeta),]
> coef1
                   coefficients     ebeta inv_ebeta
credit.policy       -0.40150505 0.6693119 1.4940717
fico                -0.30290687 0.7386679 1.3537884
log.annual.inc      -0.22851656 0.7957131 1.2567343
delinq.2yrs         -0.04734443 0.9537588 1.0484831
revol.util           0.02140484 1.0216356 0.9788226
dti                  0.02616952 1.0265149 0.9741699
pub.rec              0.02964328 1.0300870 0.9707918
days.with.cr.line    0.03980157 1.0406043 0.9609801
int.rate             0.05019803 1.0514793 0.9510411
revol.bal            0.08375515 1.0873626 0.9196564
inq.last.6mths       0.18347475 1.2013846 0.8323729
installment          0.24641762 1.2794338 0.7815957
```

Chapter 4: MODEL COMPARISON

The models are compared and analyzed here. After observing the ROC curve chart, we can see that the ROC curve is better for the logistic model we came up with than for the decision tree. Based on our analysis, we can see that the Logistic regression model has the highest AUC score of 0.67.

Though the accuracy of both models is almost the same, 83%, the precision score of the Decision tree model is 0.8457 and recall rate of 0.9859. We can see that the misclassification where the loaner has to repay is higher, to which we look into a model that has a higher recall rate. The logistic model has a recall rate of 0.9962 which is higher than the decision tree model. With the consideration of imbalanced data and misclassification, we will finalize that the logistic approach is the best fit for this data among the selected models.

Based on the odds ratio, we can choose the fields to consider for prediction. credit.policy, fico and annual income show the highest odds ratio whereas installment shows the lowest odds ratio. Based on the correlation plot Fig. 2.3.1, we can choose int.rate value as this shows a high inverse correlation with fico value. To elaborate on this, we assume the following:

- If a borrower is deemed worthy of a loan, that person is more likely to repay the full amount(nearly 1.5 times more likely than one who is deemed unworthy)
- If the borrower under consideration has a high FICO score, the odds of repayment are high

- As annual income increases, the chances of repayment increase. So, lenders can safely use annual income as an indicator.
- The chances of defaulting are higher if a borrower has to pay high amounts in installments.
- As FICO is in strong negative correlation with the interest rate, it is safe to say that chances of defaulting increase based on the increase in interest rate.

Chapter 5: CONCLUSION

More analysis and conclusions can be provided for this data. We have not considered the outlier issue in our analysis. If we take that into consideration, our predictive model will be more valid and efficient than earlier. Also, if we have larger samples, we can have more training sets which will help with the high variance problems and make our model more valid. Other methods like Random Forest, SVM(Support Vector Machine) can also be employed to create models and they can be evaluated to determine if they are a better fit. Random forest model would be used as a baseline model to better understand the dataset, because, unlike a regression model, it would be immune to multicollinearity by nature. When they decide to split, the tree will choose only one of the perfectly correlated features. Moreover, class imbalance methods that include balanced bagging can be used as a better fit than accuracy.

Loans are a common term in today's world. Lenders provide loans with the criteria of having them repaid back. However, there are scenarios where a user will be unable to pay this huge amount which leads to a huge financial loss. The possibility of the loaners knowing which category or classification of users falls in the potential defaulter's list will help them to avoid this scenario.

In this case, we have taken the data set, provided data cleaning and data exploratory analysis was performed. We dealt with imbalanced data and also categorical feature transformation along with data scaling, so as to normalize the data. The methods used to predict the model using this data set were Decision tree and Logistic Regression. We can conclude by saying that the Logistic Regression model came out as a better model with a high Area under ROC curve.

From a business perspective, we can say that loan borrowers with high income and high FICO scores are more likely to pay the loan back. In addition, those with proper financial grounds are also likely to pay the loan back. Borrowers with a low-interest rate and small installments are likely to pay back the loan fully. This means borrowers with high-interest rates and high installments will likely be on the potential defaulter's list.