

Lending Club - loan Default Prediction

BUDT758T - Data Mining Project Report



Instructor – Dr. Wei Chen

12/8/2016

**Submitted by-**

Natasha Medina

Nupur Singh

Ravi Pandey

Rong Sun

Vandna Chaturvedi

**Table of Contents**

**Executive Report** 4

**Introduction to Lending Club** 5

**Single Variable Data Exploration and Cleaning** 6

Loan Amount 6

Term 7

Interest Rate 7

Grade 8

Employment Length 9

Home Ownership 9

Annual Income 10

Verification Status 11

Issue Year 11

Loan Status 12

Purpose 14

State 15

Debt-to-income Ratio 15

Delinquency in 2 years 16

Earliest Credit Line 17

Inquiry last 6 months 17

Open Accounts 18

Revolving Balance 19

Revolving Utilization 19

Total Account 20

**Exploring Relationship (Multiple Variable) 21**

Loan Status vs Loan Amount 21

Loan Amount vs Bad Loan 21

Loan Term vs Loan Amount 22

Interest Rate vs Grade 22

Annual Income vs Grade 23

Annual Income vs Interest 23

Verification Status vs Interest 24

Home Ownership vs Interest Rate 24

**Prediction Modelling 25**

Under sample the Good loans 25

Decision Tree 25

Bagging 27

Random Forest 28

Boosting 29

Logistic Regression 30

LASSO 31

Naïve Bayes 32

**Summarizing Performance 33**

**Business Application 34**

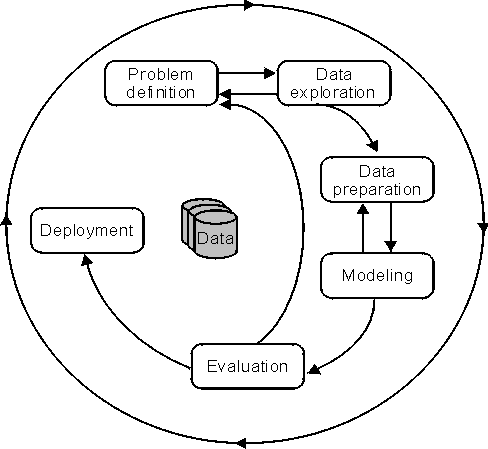
**Challenges to Successful Credit Risk Management 34**

# 

# Executive Summary

The project aims at addressing one of the most critical challenges faced by credit risk analysts and various bankers day-to-day i.e. reducing default/bad loans. Bad loans are a huge overhead to all the financial institutions. According to a survey in 2015-2016, more than 16% of all the loans in United States go default, which not only reduces the profit margin for financial institution but sometimes causes huge losses for them.

In this project we have tried to handle the issue of one such financial organization, Lending Club. Lending Club is the world’s largest online marketplace for investors and borrowers. Any US citizen can become an investor or can become a borrower in this marketplace. However, becoming an investor entails the dilemma of extending loan to a borrower who might default.

In this project, we solve this dilemma of the investors. **We have come out with predictive models that can predict loan defaults with an accuracy as high as 75%.** We started this project by collecting historical data from Lending club. The website provided us data from 2007 to 2016. Further, we started exploring and processing the data. Various charts, plots and summary statistics provided us useful insights from the data. After processing and visualizing the data, we moved on to build predictive models. **We tried 7 different models, of which Random Forest proved the most accuarte in predicting defaullts.**

Along with performance measures, we also came out with various insights from the dataset. We found out that interest rate, loan amount, annual income and debt-to-income ratio were some of the most important variables. Also, grades given by lending club to each borrower/loan are a direct impact of the interest rate charged from the borrower. We should also note that FICO score, one of other criteria to measure person’s creditworthiness was not used in the analysis. In future, we would like to implement this model in real world scenario and note down its performance with real loans.

# Introduction to Lending Club

# Lending club, known as the world’s largest marketplace, provides a platform to connect borrowers and investors. As a borrower, one not only gets loan at a lower rate of interest but also goes through an easier and faster process than traditional bank loans. As an investor, one can earn attractive, risk-adjusted returns by customizing his investment portfolio. The entire process is online, using technology to improve the efficiency and lower the cost, which earned Lending Club the highest satisfaction ratings in the financial services industry.

Borrowers start with filling an online application, putting his personal information such as amount of dollars he wants to borrow, purpose, his annual income and so on to enable lending club evaluate the borrower’s credibility and the associated loan rate. Once an application has been assigned a grade from A to G, from the lowest risk to highest risk level, investors are able to customize their investment portfolio from a great range of investment selection. All in all, a borrower with higher credibility and thus a lower interest rate is assigned a higher grade, and the associated return for investors who put money in is lower due to the lower risk.

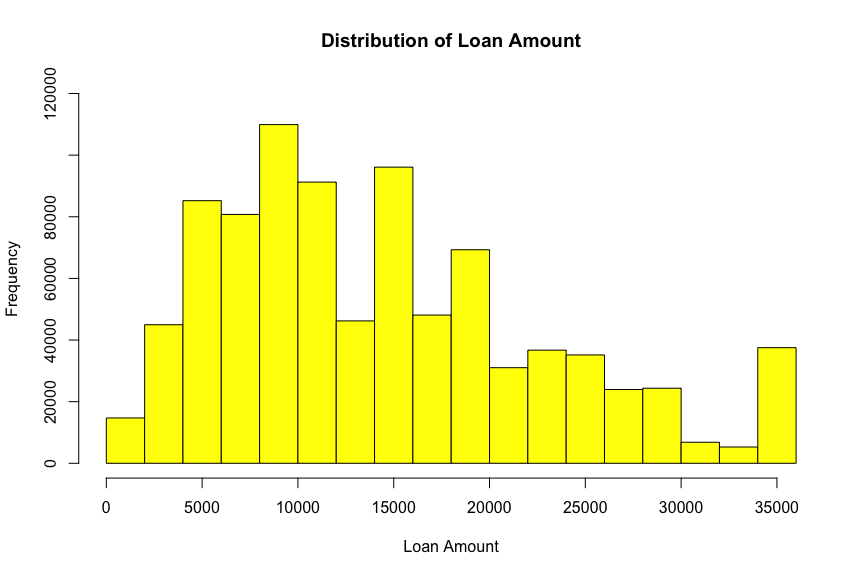
Lending club possesses some features which aims to provide best friendly services to both borrowers and investors. For example, a borrower can check his rate fast and free and he can watch as people invest in his loan. Additionally, fast funding go directly to the connected bank account which is convenient to the borrower. As for the investors, opening an account and transferring funds are quickly and easily as well. Furthermore, investors are allowed to build customized portfolio by investing in a range of loans as little as $25, and they receive monthly payments of principal and interest as borrowers repay their loans. As the largest lending platform, Lending Club leverages online data and technology to quickly assess risk, determine a credit rating and assign appropriate interest rates to each applicant. Our project tried to imitate and find the best model to predict an applicant’s credit level based on the given information.

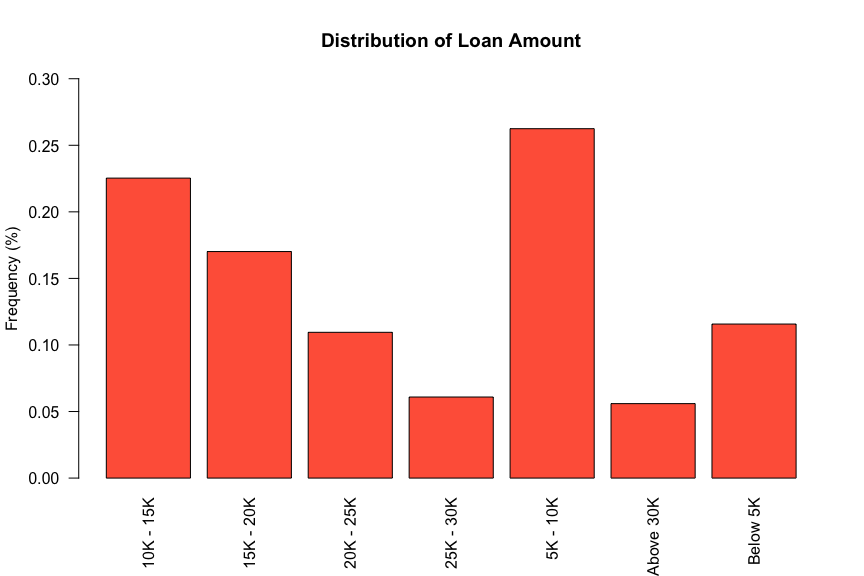
# Single Variable Data Exploration and Cleaning

## **Loan Amount**

Loan amount is the amount of loan that has been extended to the borrower. In our dataset, the amount ranged from $500 - $35,000 and the maximum number of loans were borrowed in the $7,500-$10,000 range.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 500 8000 13000 14760 20000 35000

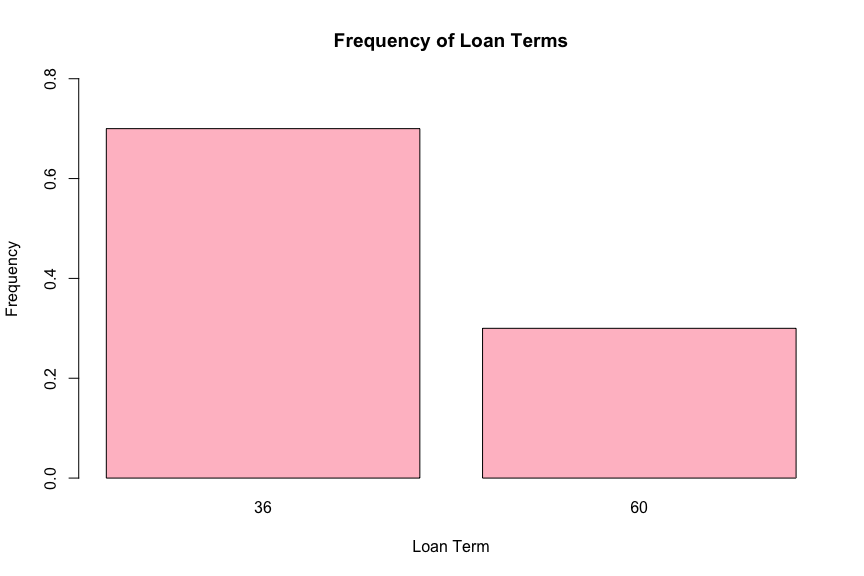




## **Term**

Term is the period for which the loan is extended and is measured by the number of payments on the loan. Values are in months which are either 36 or 60. It was seen that most of the loans, 70%, were taken on 36-month payment plan.

##   
## 36 60   
## 621125 266254

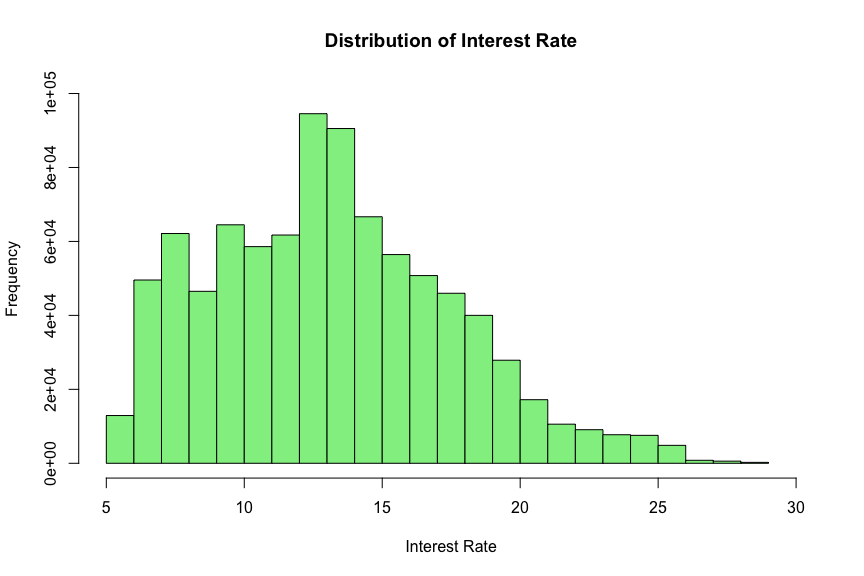


## 

## **Interest Rate**

Interest Rate is the rate of interest on the loan. As can be seen from the histogram, it is somewhat right skewed and even though the interest rate ranges from 5.32-28.9, a big part of the loans were extended for an interest rate up to 20%, with the average rate being 13.25%.

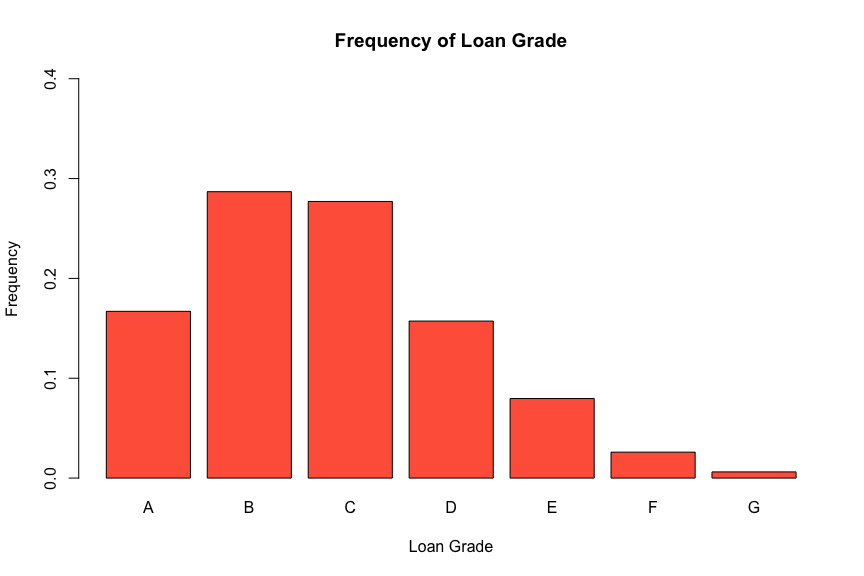
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 5.32 9.99 12.99 13.25 16.20 28.99



## **Grade**

Grade field is the grade assigned to each loan by Lending Club. It reflects how likely the loan is to be paid off and is determined based on the creditworthiness of the borrower. The grades in the dataset range from A-G, B and C being the most assigned and very few G grade loans extended.

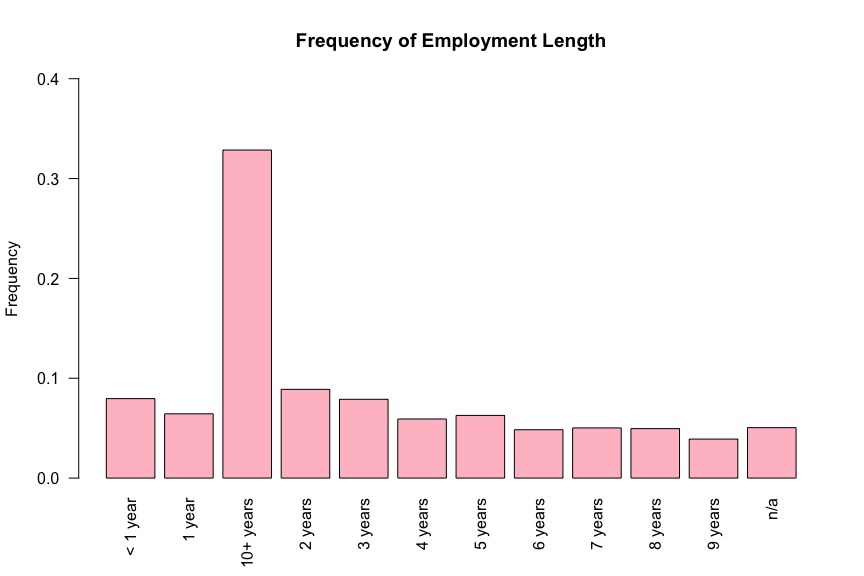
##   
## A B C D E F G   
## 148202 254535 245860 139542 70705 23046 5489



## **Employment Length**

Employment length is the length in years for which the borrower has been employed. Possible values are between 0 and 10 where 0 being less than one year and 10 being ten or more years. N/a is the value assigned to the borrower who has never been in employment so far. The graph clearly shows that a vast majority of the loans have been extended to people with 10 or more years in employment.

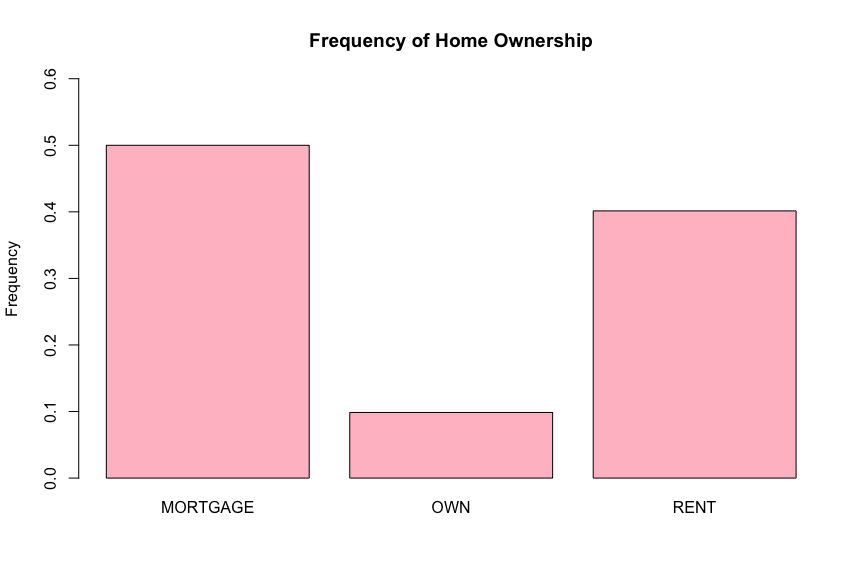
##   
## < 1 year 1 year 10+ years 2 years 3 years 4 years 5 years   
## 70605 57095 291569 78870 70026 52529 55704  
## 6 years 7 years 8 years 9 years n/a   
## 42950 44594 43955 34657 44825



## **Home Ownership**

The home ownership status provided by the borrower during registration. Our values include RENT, OWN, MORTGAGE, OTHER, however, as is evident from the visualization, none of the applicants selected ‘Other’. Also, close to 50% of the loans were extended to people having an existing mortgage and those owning a house formed the lowest part, being 10%.

##   
## MORTGAGE OWN RENT   
## 443560 87470 356117

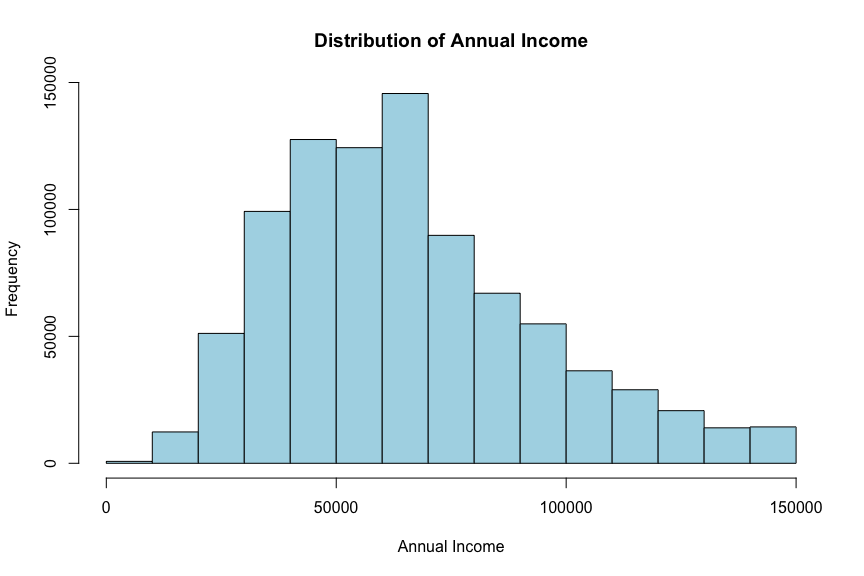


## **Annual Income**

The annual income is what was provided by the borrower during registration. In our dataset, the annual income ranges from anywhere between $0- $150,000 and as can be seen in the plot, it is somewhat right skewed with the majority of borrowers making within the $100,000 range, the average being around $67,000.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 45000 65000 75030 90000 9500000

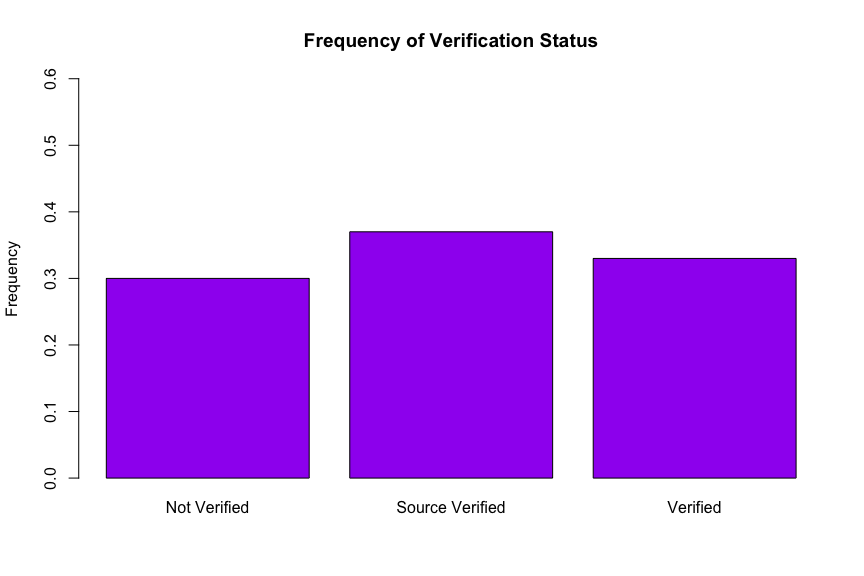
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 45000 65000 67030 84000 150000



## **Verification Status**

Indicates whether the co-borrowers' joint income has been verified by Lending Club. It is classified into ‘Verified’, ‘Not verified’ or ‘Source verified’ and from the frequency plot, it can be seen that the cases belonging to each category are uniformly distributed with 37% being source verified and 30% not verified.

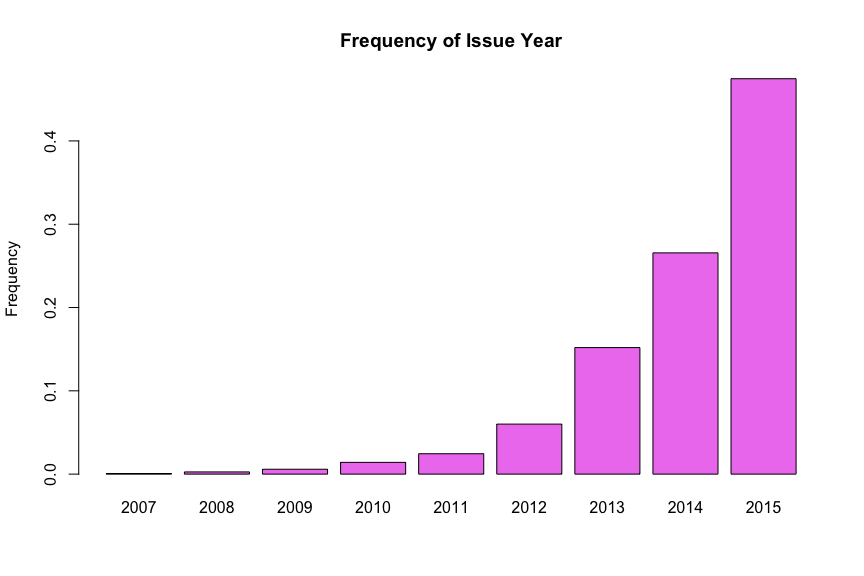
##   
## Not Verified Source Verified Verified   
## 266633 329537 290977



## **Issue Year**

Lending Club started in 2007 started issuing 596 loans. As the year’s progress, Lending Club increased the number of loans per year by thousands. During the year of 2012 there was a huge number of loans issued of 53,279 loans compare to last year of 2015 issued 421,094 loans. The graph displays as the years progressed which is shown in the graph Lending Club has increased the number of loans over the years. At the year of 2012, Lending Club commenced to issue more loans all the way up to 2015. The graph displays there is an increase trend of issue loans for this year.

##   
## 2007 2008 2009 2010 2011 2012 2013 2014 2015   
## 596 2345 5196 12534 21720 53279 134755 235628 421094

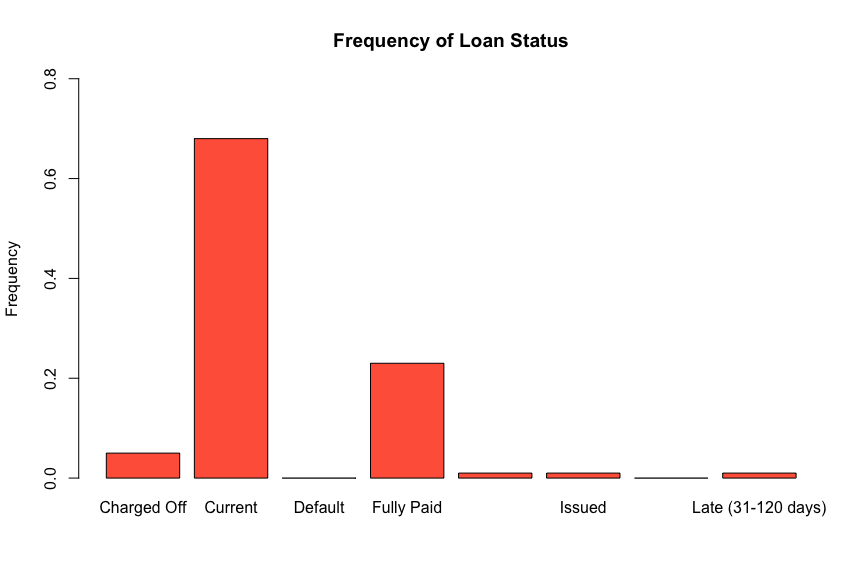


## **Loan Status**

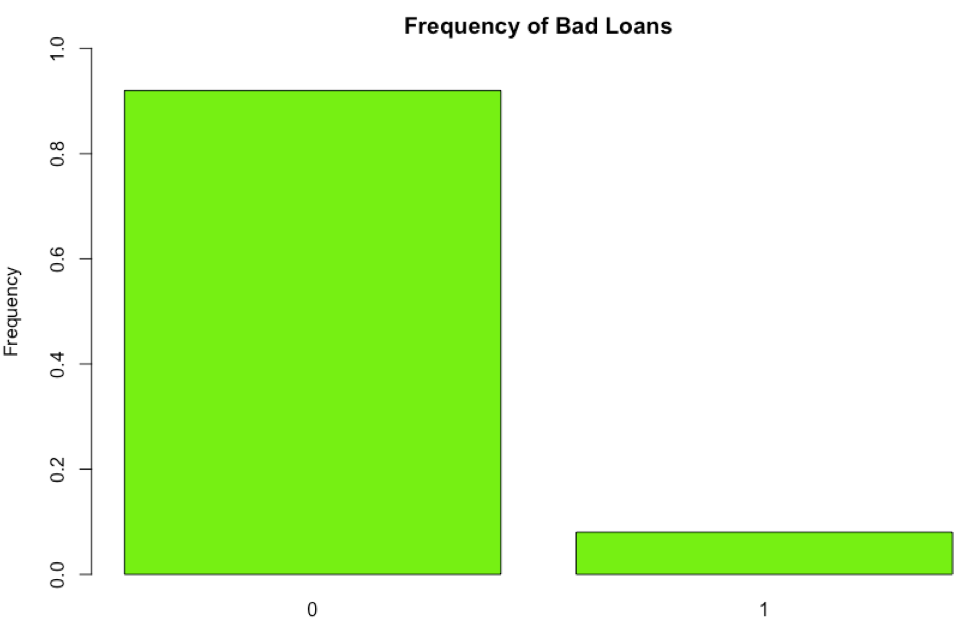
The term loan status means the current status of the borrower. There are 8 different categories a borrower can be placed under which are the following: charged off, current, default, fully paid, in grace period, issued, last (16-30 days), late (31-120 days). The largest category borrowers are placed in our current dataset, second largest which is 207,573 loans and third largest is charged off. Charged off status implies that loans for which are no longer a reasonable expectation of further payments. Generally, Charge Off occurs no later than 30 days after the Default status is reached.

As you can see in the diagram labelled “Frequency, distribution is slight skewed to the right. 70 percent of borrowers which is majority are labelled as current status. 30 percent of borrowers are labelled as fully paid under loan status. The second top category which means loan has been fully repaid, either at the expiration of the 3 or 5-year term as a result of prepayment. 10 percent of borrowers are labelled as charged off which implies that the borrower who are labelled as status, means that no longer a reasonable expectation of further payments according to (Lending club website.)

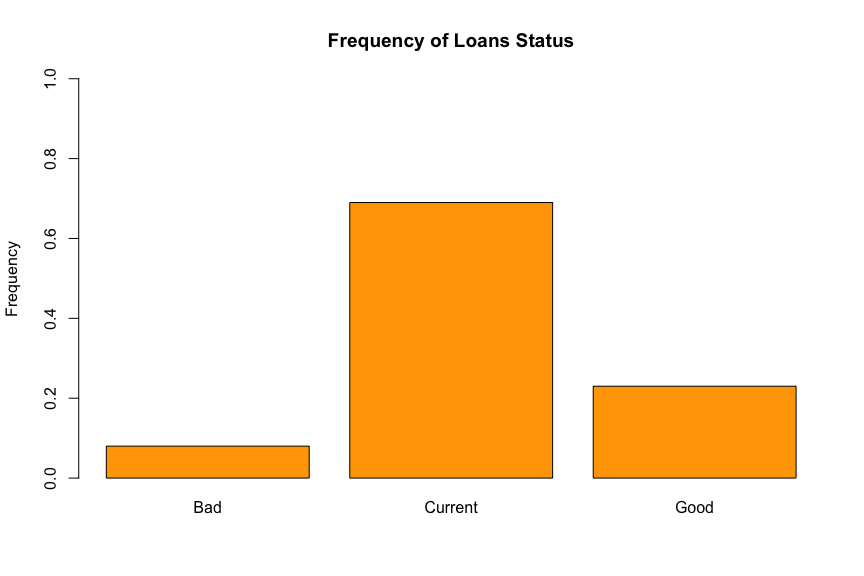
##   
## Charged Off Current Default   
## 45214 601774 1219   
## Fully Paid In Grace Period Issued   
## 207573 6253 8460   
## Late (16-30 days) Late (31-120 days)   
## 2357 11591



For the graph titled, “Frequency of Bad Loans,” we found out that the frequency of a borrower getting a bad loan is 0.1, and a good one is .9.



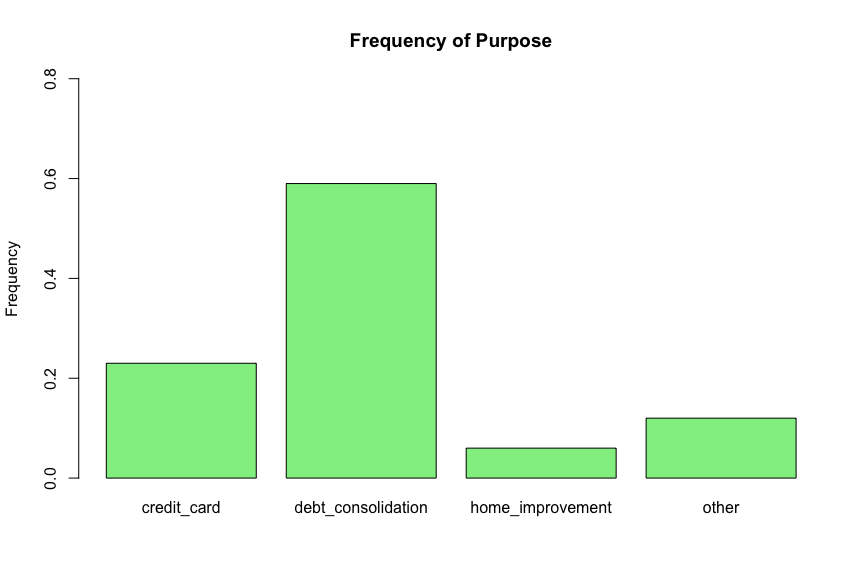
For graph titled, “Loan Status,” visualized the frequency of a borrower being labelled as current, bad or good to see if he or she will default in their loans. The chart displays 70 percent of borrowers are labelled as current under loan status. 30 percent of borrowers are labelled as good under loan status, and 10 percent of borrowers are labelled as bad.



## **Purpose**

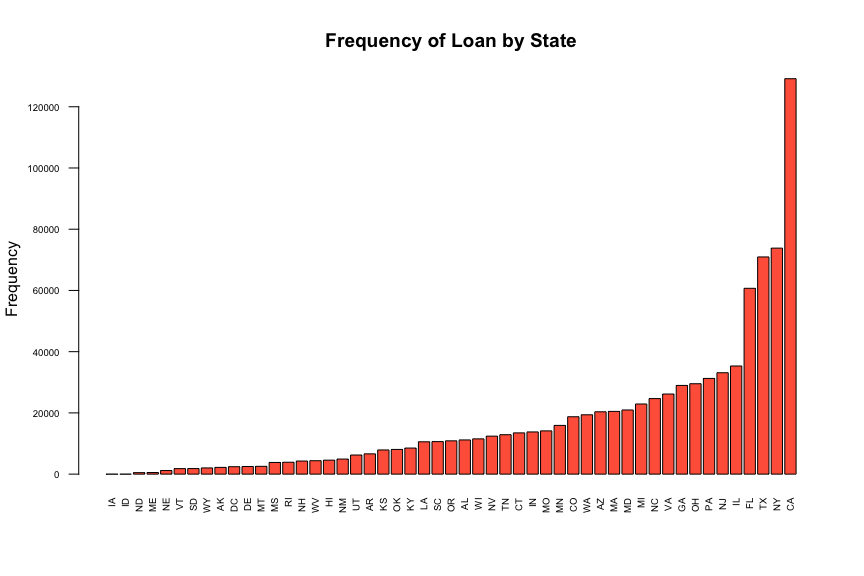
There are four categories that a borrower fills out in application describes purpose why he or she wants to borrow a loan the following are: credit card, debt consolidation, home improvement, and others. The highest category that a borrower asks to borrow a loan is debt consolidation .60 percent of people borrow money for debt consolidation. 23 percent of borrowers seek for a loan from Lending Club for credit card. 25 percent of borrowers seek loan for other type of reasons from Lending Club.

##   
## credit\_card debt\_consolidation home\_improvement   
## 205776 523051 51602   
## other   
## 104012



## **State**

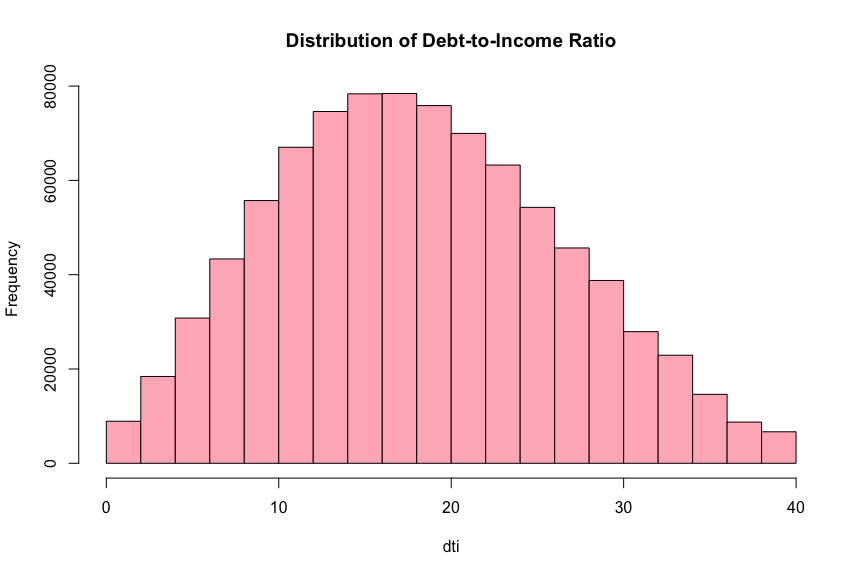
We wanted to visualize from what states the borrowers apply from, we could conclude that California is twice as much compare to the rest of the states where the applicants are from. The graph is gradually increasing in ascending order. California has 120,000 applicants from that state compare to New York which is half of the size compare to California. We can draw a conclusion that for the top four states California, New York, Texas, and Florida are metropolitan’s states with more jobs to payback the loans they borrowed from Lending Club.



## **Debt-to-income Ratio**

Debt to income ratio, which is the percentage of a consumer’s monthly gross income that goes toward paying debts (excluding household debt). Lending Club uses DTI as part of its “risk modifiers” that help to calculate a “loan grade” ranging from A to G. The loan grade is mostly based on a consumer’s FICO scores, but is modified to take into accounts the DTI and the loan amounts. The distribution of debt to Income seems to be normally distributed bell shape curve. Large number of borrowers in Lending Club DTI score majority falls from 10 to 23.96 even though the maximum is 39.99.

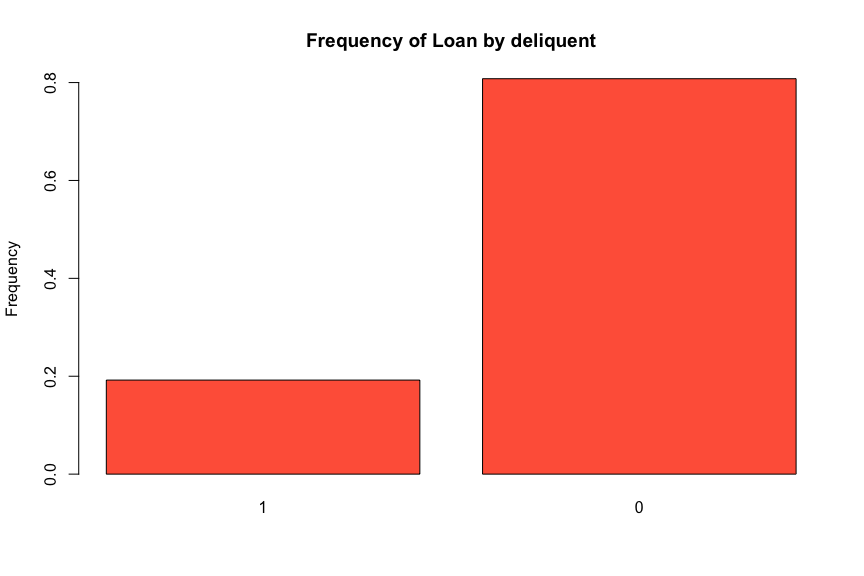
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 11.92 17.66 18.14 23.96 39.99



## **Delinquency in 2 years**

Delinquency in 2 years means that the borrower has been due for a payment more than 30 days in the past two years. The frequency of the bar graph shows number of borrowers of loan by delinquent in two years whether they are delinquent or not. We can infer that 81 percent of borrowers are not delinquent in the last 2 years and 19 percent of the borrowers have been delinquent in the last 2 years.

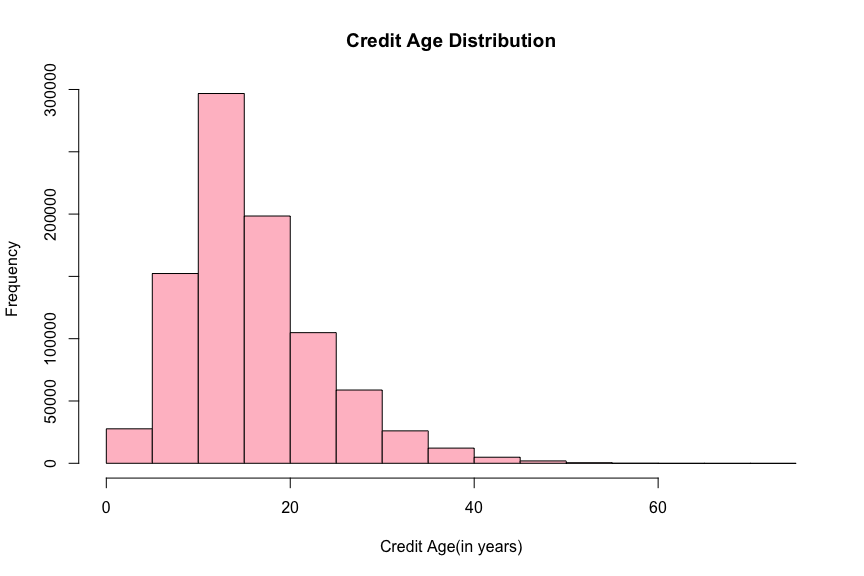
##   
## 0 1   
## 714432 169919



## **Earliest Credit Line**

Earliest credit line is the month the borrower's earliest reported credit line was opened. And from the frequency histogram, we see a large number of people fall in the category with 10-15 month of credit age. Even though the maximum credit age is 71, around 90% of total people has first credit line opened within 30 months.

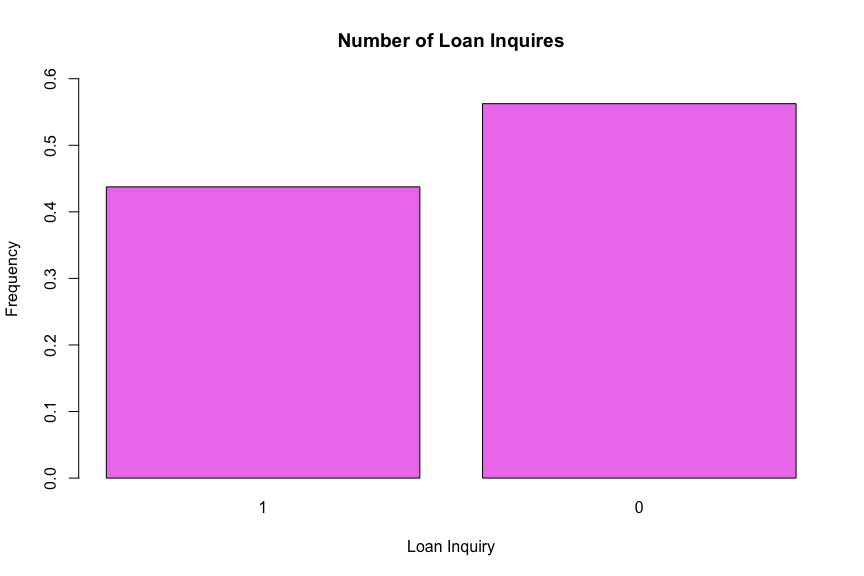
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 3.00 11.00 15.00 16.34 20.00 71.00



## **Inquiry last 6 months**

The frequency bar shows the number of inquiries in past 6 months (excluding auto and mortgage inquiries). Number of people who have 1 loan inquiry accounts for 45% of total people, and almost 55% of total observations have 0 loan inquiry in the past 6 months.

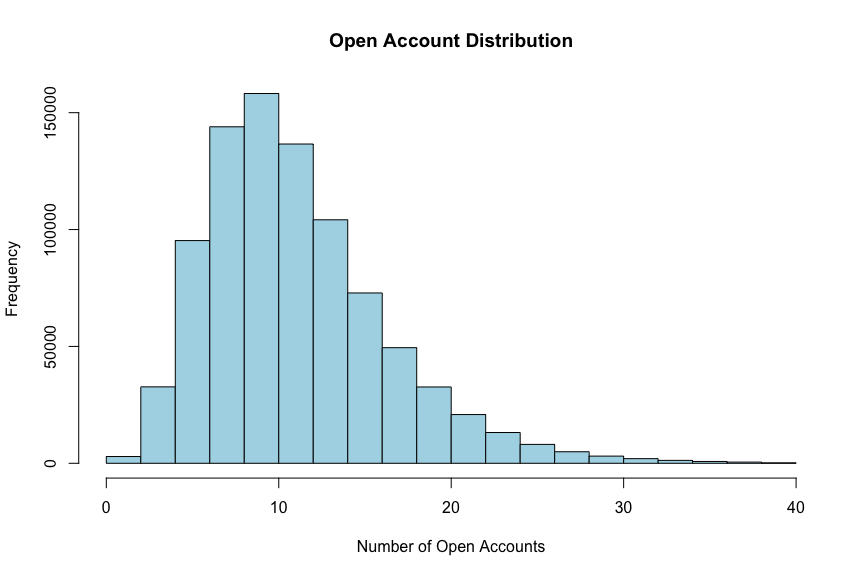
##   
## 0 1   
## 497445 386906



## **Open Accounts**

Open accounts measure the number of open credit lines in the borrower's credit file. The distribution is approximately normal with a mean of 11.52 and standard deviation of 5.16, but it is highly right-skewed as well. In other words, number of open credit lines beyond 30 are considered as outliers which occur rarely.

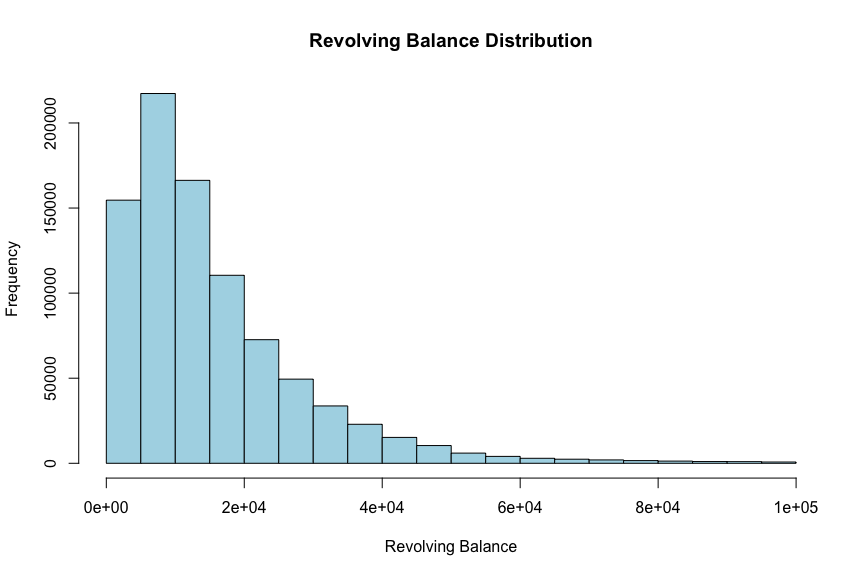
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 8.00 11.00 11.52 14.00 39.00



## **Revolving Balance**

Revolving balance is the total credit revolving balance of a person. The distribution is highly right skewed which means many outliers occur in the data. The average for revolving balance is 15520, and more than half of the observations have less than 20,000 of total credits. The range of revolving range is from 0 to 99980.

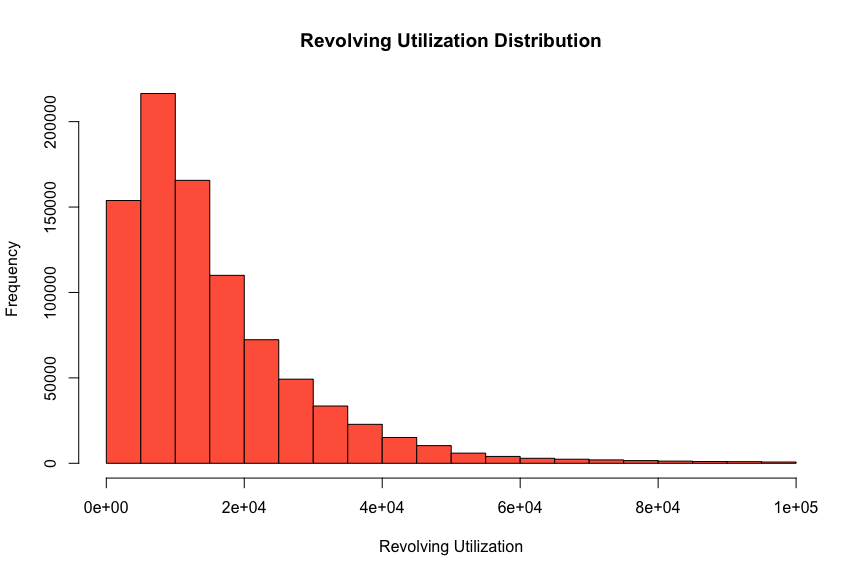
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0 6410 11770 15520 20470 99980



## **Revolving Utilization**

Revolving utilization stands for the revolving line utilization rate, or the amount of credit the borrower is using relative to all available revolving credit. The distribution is highly right-skewed, and the frequency decreases as the revolving utilization decreases after it reaches the peak. The range of revolving utilization is from 0 to 100 and the mean is 54.82.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 0.00 37.60 55.80 54.82 73.20 100.00

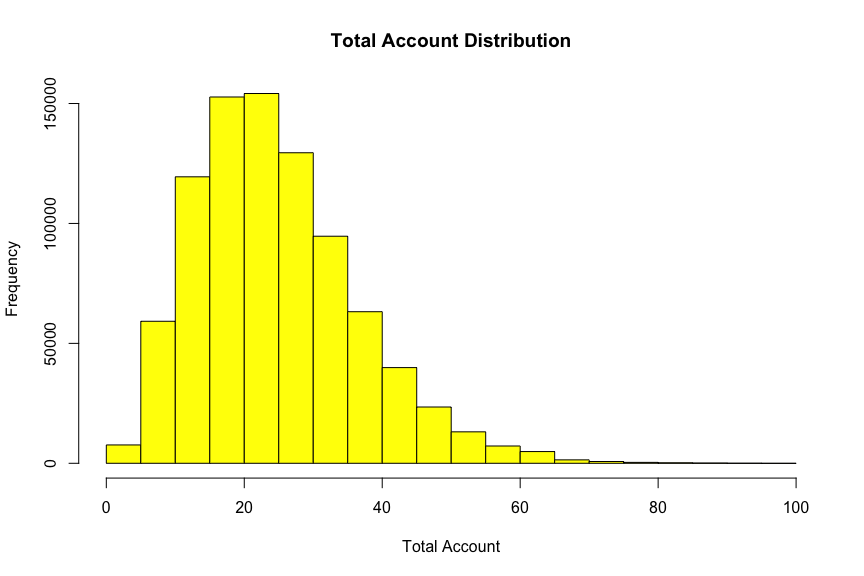


## 

## **Total Account**

Total account is the total number of credit lines currently in the borrower's credit file. The distribution looks normal with a little bit right skewness. The most frequently total account is between 15 and 25 and the total range is from 2 to 100. The average total account is about 25.

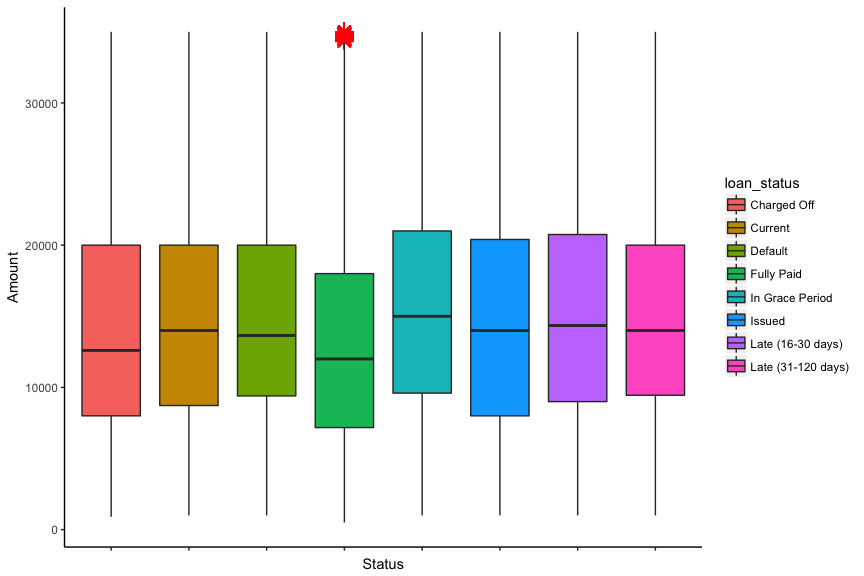
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 2.0 17.0 24.0 25.2 32.0 100.0



# Exploring Relationship (Multiple Variable)

## **Loan Status vs Loan Amount**

The relationship shown in this graph is between Loan Status and Loan amount. The lowest average loan amount has the loan status of Fully paid, which suggests that people are most likely to pay off the smallest loan amount. Similarly, the highest average loan amount has the status of loan as Grace period, so the loan amounts which are higher are in grace period and have not been fully paid.

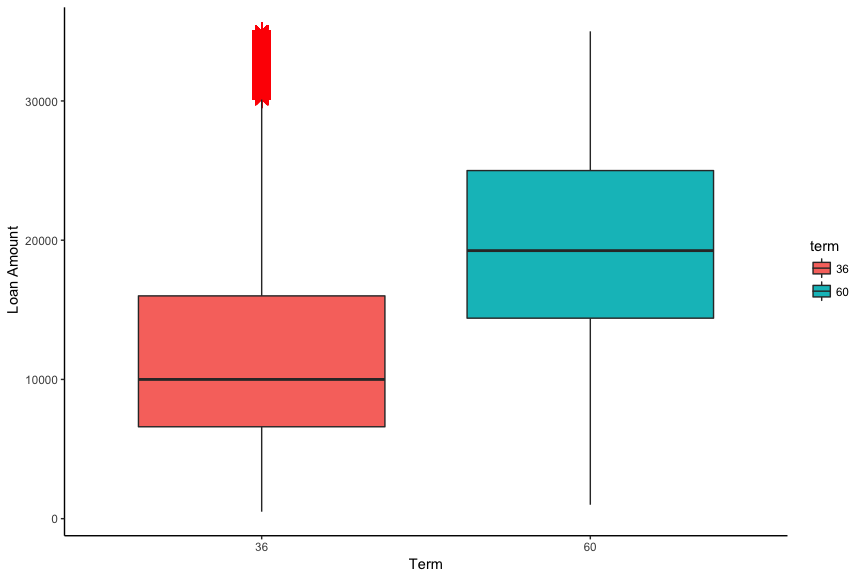


## **Loan Amount vs Bad Loan**

The relationship between Loan amount and Loan status shows that the loan amount does not affect whether that loan is good or bad, as the average loan amount is the same for both.

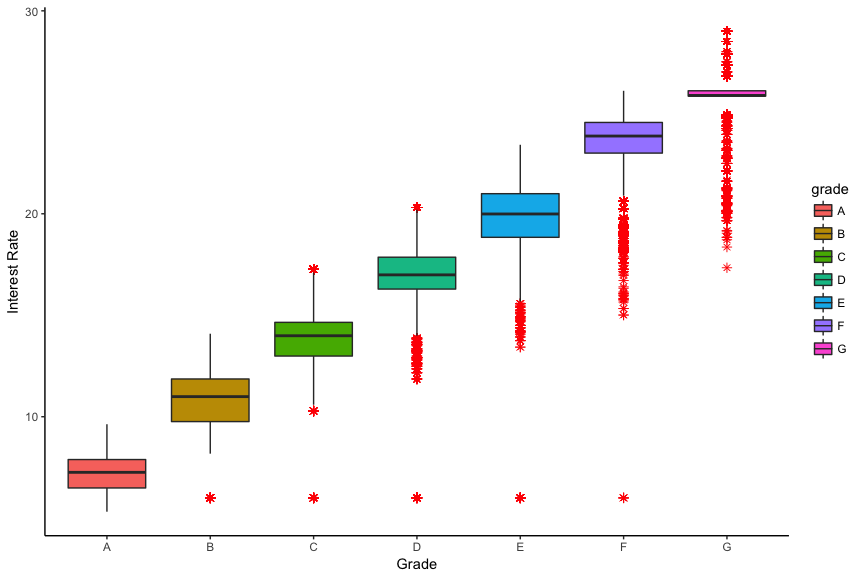
## **Loan Term vs Loan Amount**

The relationship between Loan term and Loan amount suggests that if the term of loan is less that is 36 months, the average loan amount will be less than the average loan amount if the term of loan is more, that is 60 months.



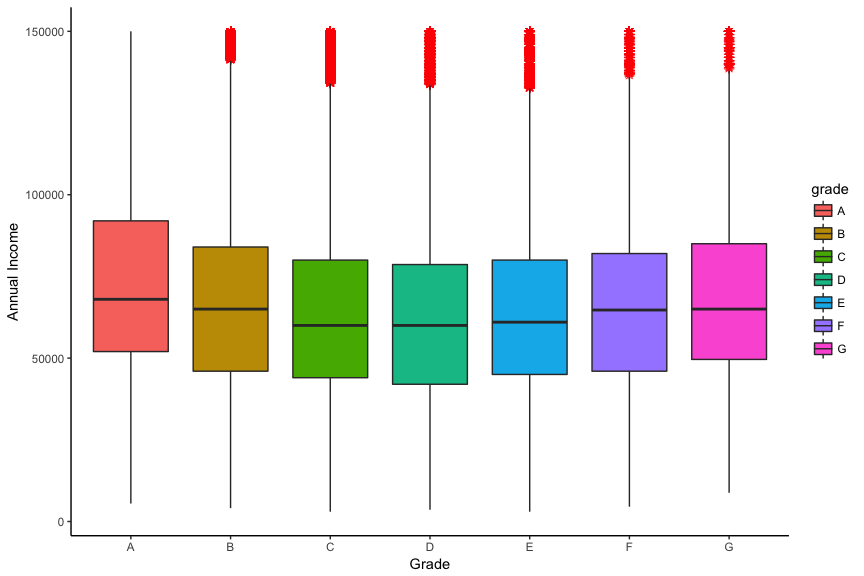
## **Interest Rate vs Grade**

The relationship between Interest rate and Grade suggests that the Grade is the highest for loans with a lower rate of interest and lowest for the highest interest rate loans, grade G being assigned to loans with interest between 20 -30 % and Grade A being assigned to loans with interest less than 10 %.



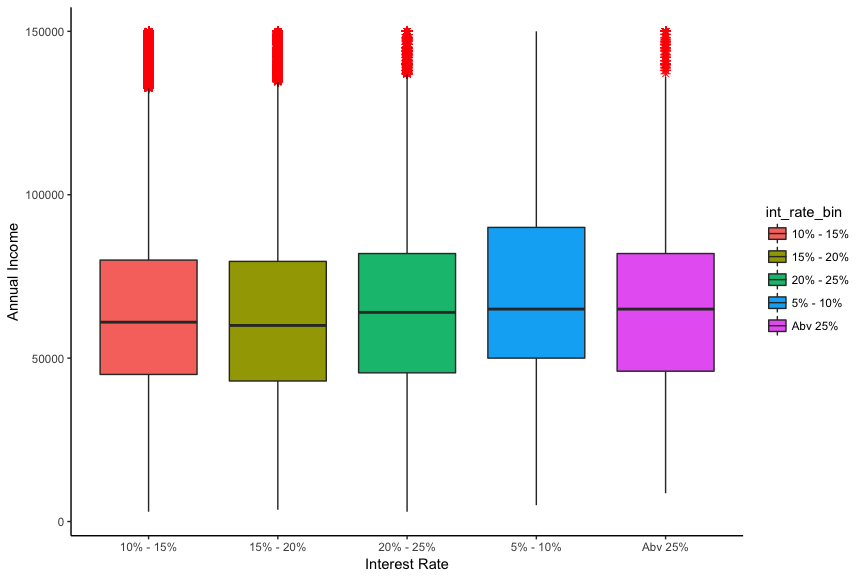
## **Annual Income vs Grade**

The relationship between annual income and Grade suggests a high income corresponds to grade A, as a borrower with higher income is more likely to repay the loan. Grade D is assigned to a borrower with a median average income. Grade G is assigned to a borrower with higher income that the rest but slightly lower income than the grade A contender.



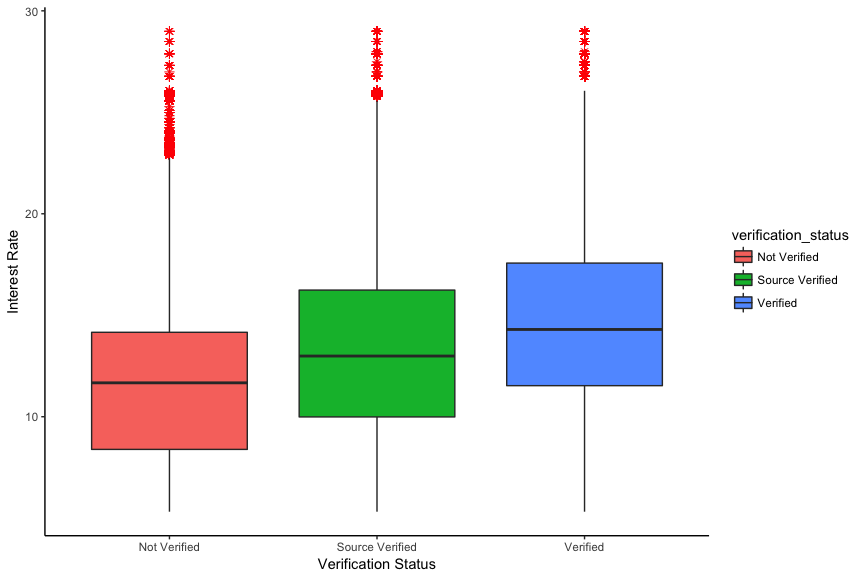
## **Annual Income vs Interest**

The relationship between annual income and interest suggests that the lowest interest rate of 5- 10 % is given to the borrowers with have an average highest average annual income.



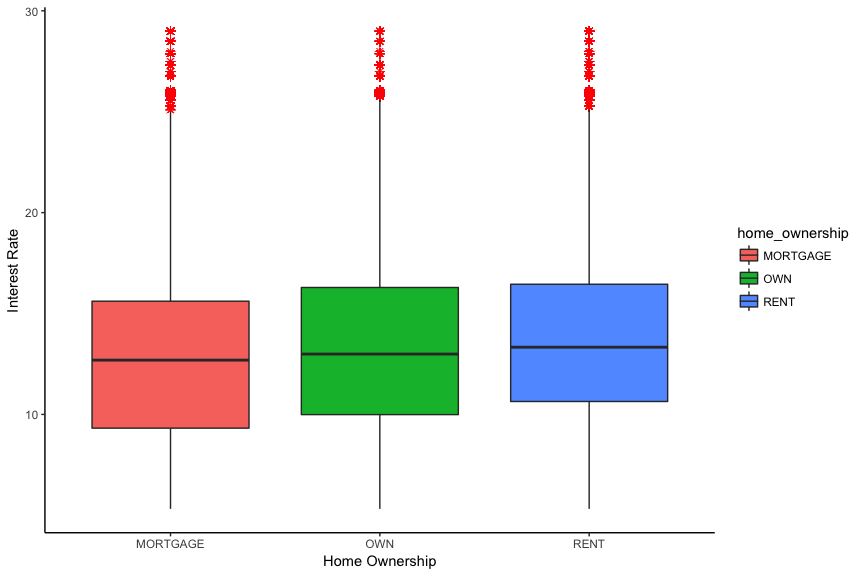
## **Verification Status vs Interest**

The relationship between verification state and interest rate suggests that the borrowers that have a verified income source are more likely to a high interest rate. However, the lowest interest rate is of the borrowers with not verified income source because they are mostly declined the loan or the loan lent must be very less, therefore a low interest rate.



## **Home Ownership vs Interest Rate**

The relationship between Home ownership status and interest rate suggests that the borrowers with a status of Mortgage are offered the lowest interest rate and the ones on rent are offered the highest interest rate.



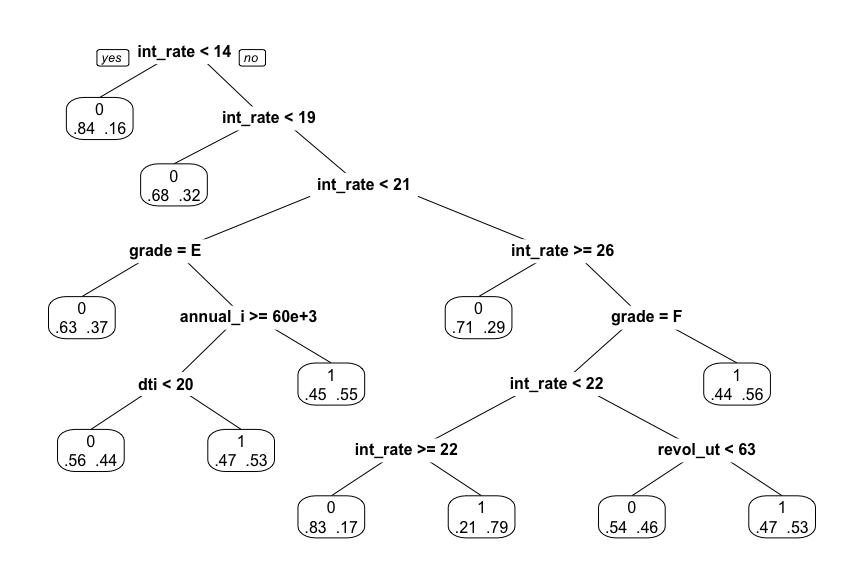
# Modelling

# Under sample the Good loans

The data available to us is highly unbalanced with approximately 1% bad vs 99% good loans. In such scenario, our model may not give us required results. One of the methods to solve this problem is to under sample the good loans. After under sampling, we will have approx. 75%-25% good vs bad loans, which will improve our model prediction.

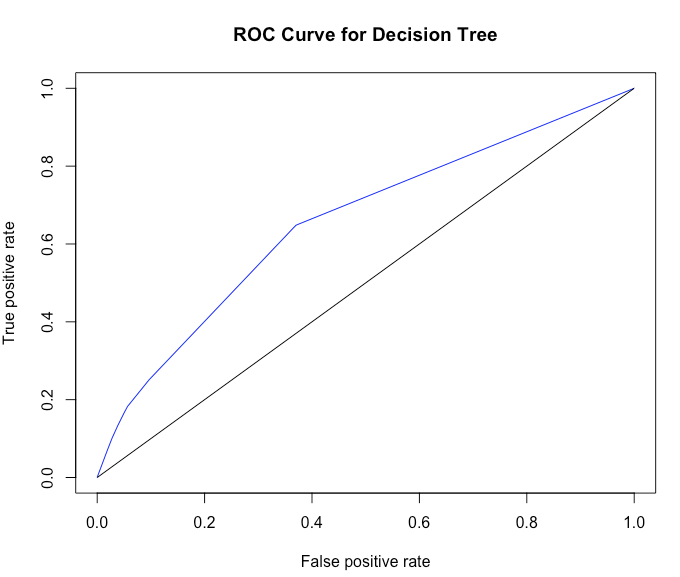
# Decision Tree

Tree model is used to predict either a qualitative or quantitative result based on what region the observation falls into. For regression tree, we built the model and use the best split to minimize RSS each step, while for classification tree, we use misclassification error rate instead. Once tree model has been developed, we can just go through each node and the associated branch for predicted value.



|  |  |
| --- | --- |
| ## Confusion Matrix and Statistics (cut-off = 20%)  ##  ## Actual ## Prediction 0 1 ## 0 37338 6953 ## 1 21965 12814 ##  ## Accuracy : 0.6343 ## Sensitivity : 0.6483  ## Specificity : 0.6296 | Confusion Matrix and Statistics (cut-off = 25%) ##  ## Actual ## Prediction 0 1 ## 0 37338 6953 ## 1 21965 12814 ##  ## Accuracy : 0.6343  ## Sensitivity : 0.6483  ## Specificity : 0.6296 |

|  |  |
| --- | --- |
| Confusion Matrix and Statistics (cut-off = 30%) ##  ## Actual ## Prediction 0 1 ## 0 37416 6984 ## 1 21887 12783 ##  ## Accuracy : 0.6349   Sensitivity : 0.6467  ## Specificity : 0.6309 | Confusion Matrix and Statistics (cut-off = 50%) ##  ## Actual ## Prediction 0 1 ## 0 56834 16949 ## 1 2469 2818 ##  ## Accuracy : 0.7544   Sensitivity : 0.1425  ## Specificity : 0.9583 |



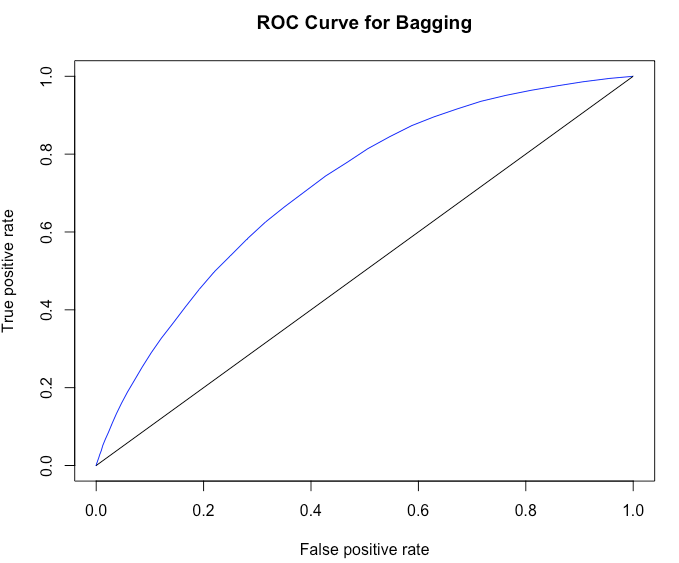
## [1] "Area Under Curve for decision tree is : 0.65631"

**Bagging**

Bagging is the method of taking repeated sample from a single training dataset to reduce the variance and hence increase the prediction accuracy. To start with, we need to create a finite number of datasets using sample with replacement. Then, we create one classifier for each dataset. Lastly, we take average for numeric prediction but use majority vote for categorical outcome prediction.

|  |  |
| --- | --- |
| ## Confusion Matrix and Statistics (cut-off = 20%)  ##  ## Actual ## Prediction 0 1 ## 0 29313 3678 ## 1 29990 16089 ##  ## Accuracy : 0.5742 ## Sensitivity : 0.8139  ## Specificity : 0.4943 | ## Confusion Matrix and Statistics (cut-off = 25%) ##  ## Actual ## Prediction 0 1 ## 0 33963 5062 ## 1 25340 14705 ##  ## Accuracy : 0.6155  ## Sensitivity : 0.7493  ## Specificity : 0.5727 |

|  |  |
| --- | --- |
| Confusion Matrix and Statistics (cut-off = 30%) ##  ## Actual ## Prediction 0 1 ## 0 40606 7401 ## 1 18697 12366 ##  ## Accuracy : 0.6699   Sensitivity : 0.6256  ## Specificity : 0.6847 | Confusion Matrix and Statistics (cut-off = 50%) ##  ## Actual ## Prediction 0 1 ## 0 55102 15455 ## 1 4201 4312 ##  ## Accuracy : 0.7514   Sensitivity : 0.2181  ## Specificity : 0.9291 |



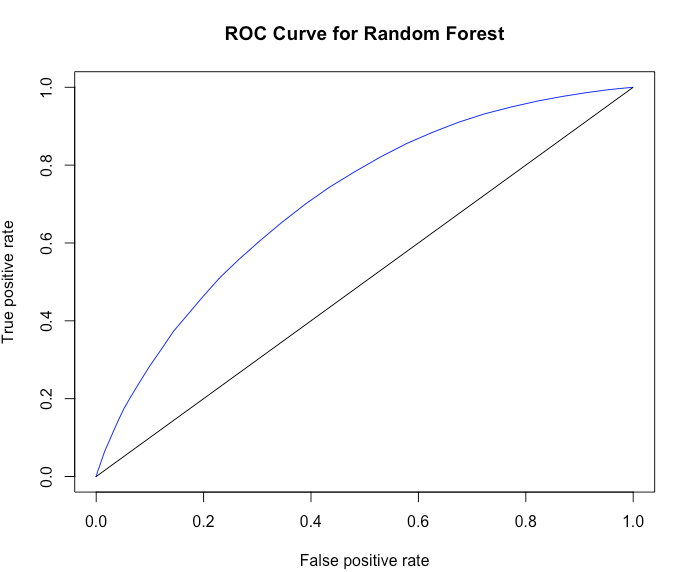
## [1] "Area Under Curve for decision tree is : 0.71656"

**Random Forest**

It is a method involving averaging multiple deep decision trees, trained on different parts of the same training set, with the goal of overcoming over-fitting problem of individual decision tree. It is an ensemble learning method that operates by constructing a lot of decision trees at training time and outputting the class that is the mode of the classes output by individual trees For many data sets, it produces a highly accurate classifier. It has an effective method for estimating missing data and maintains accuracy when a large proportion of the data are missing.

|  |  |
| --- | --- |
| ## Confusion Matrix and Statistics (cut-off = 20%)  ##  ## Actual ## Prediction 0 1 ## 0 27899 3537 ## 1 31404 16230 ##  ## Accuracy : 0.5581 ## Sensitivity : 0.8211  ## Specificity : 0.4704 | Confusion Matrix and Statistics (cut-off = 25%) ##  ## Actual ## Prediction 0 1 ## 0 33500 5061 ## 1 25803 14706 ##  ## Accuracy : 0.6097  ## Sensitivity : 0.7440  ## Specificity : 0.5649 |

|  |  |
| --- | --- |
| Confusion Matrix and Statistics (cut-off = 30%) ##  ## Actual ## Prediction 0 1 ## 0 41230 7797 ## 1 18073 11970 ##  ## Accuracy : 0.6728   Sensitivity : 0.6056  ## Specificity : 0.6952 | Confusion Matrix and Statistics (cut-off = 50%) ##  ## Actual ## Prediction 0 1 ## 0 56257 16332 ## 1 3046 3435 ##  ## Accuracy : 0.7549   Sensitivity : 0.1737  ## Specificity : 0.9486 |



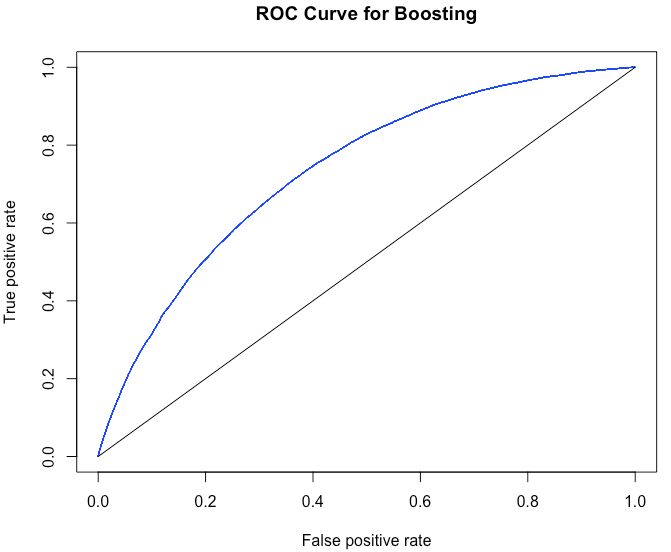
## [1] "Area Under Curve for decision tree is : 0.71229"

**Boosting**

It is a machine learning ensemble meta-algorithm for primarily reducing bias, and also variance in supervised learning, and a family of machine learning algorithms which convert weak learners to strong ones. The main advantage of boosting is the speed. As opposed to random forests, in boosting, the growth of a particular tree takes into account the other trees that have already been grown and thus smaller trees are sufficient, which aids in interpretability.

|  |  |
| --- | --- |
| ## Confusion Matrix and Statistics (cut-off = 20%)  ##  ## Actual ## Prediction 0 1 ## 0 30708 3663 ## 1 28595 16104 ##  ## Accuracy : 0.5920 ## Sensitivity : 0.8147  ## Specificity : 0.5178 | Confusion Matrix and Statistics (cut-off = 25%) ##  ## Actual ## Prediction 0 1 ## 0 37965 5794 ## 1 21338 13973 ##  ## Accuracy : 0.6569  ## Sensitivity : 0.7069  ## Specificity : 0.6402 |

|  |  |
| --- | --- |
| Confusion Matrix and Statistics (cut-off = 30%) ##  ## Actual ## Prediction 0 1 ## 0 44123 8166 ## 1 15180 11601 ##  ## Accuracy : 0.7047   Sensitivity : 0.5869  ## Specificity : 0.7440 | Confusion Matrix and Statistics (cut-off = 50%) ##  ## Actual ## Prediction 0 1 ## 0 56711 16382 ## 1 2592 3385 ##  ## Accuracy : 0.7600   Sensitivity : 0.1712  ## Specificity : 0.9562 |



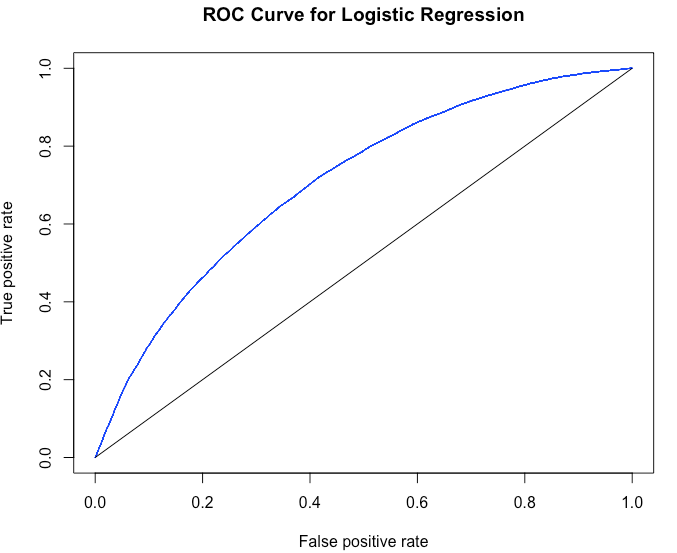
## [1] "Area Under Curve for decision tree is : 0.73566"

**Logistic Regression**

Logistic Regression is a form of regression that allows the prediction of discrete variables by a mix of continuous and discrete predicators. Logistic regression is often used because the relationship between discrete variable(DV) and predicator is non-linear. Instead of using Y (or p) as the dependent variable, we use a function of it, which is called logit. The key advantage of logits maps any value of the dependent variables into a probability [0,1]. The logit, it turns out, can be modelled as a linear function of predicators. Once the logit has been predicted, it can be mapped back to a probability p.

|  |  |
| --- | --- |
| ## Confusion Matrix and Statistics (cut-off = 20%)  ##  ## Actual ## Prediction 0 1 ## 0 29369 4085 ## 1 29934 15682 ##  ## Accuracy : 0.5698 ## Sensitivity : 0.7933  ## Specificity : 0.4952 | Confusion Matrix and Statistics (cut-off = 25%) ##  ## Actual ## Prediction 0 1 ## 0 37762 6655 ## 1 21541 13112 ##  ## Accuracy : 0.6434  ## Sensitivity : 0.6633  ## Specificity : 0.6368 |

|  |  |
| --- | --- |
| Confusion Matrix and Statistics (cut-off = 30%) ##  ## Actual ## Prediction 0 1 ## 0 44528 9294 ## 1 14775 10473 ##  ## Accuracy : 0.6956   Sensitivity : 0.5298  ## Specificity : 0.7509 | Confusion Matrix and Statistics (cut-off = 50%) ##  ## Actual ## Prediction 0 1 ## 0 56853 17051 ## 1 2450 2716 ##  ## Accuracy : 0.7534   Sensitivity : 0.1374  ## Specificity : 0.9586 |



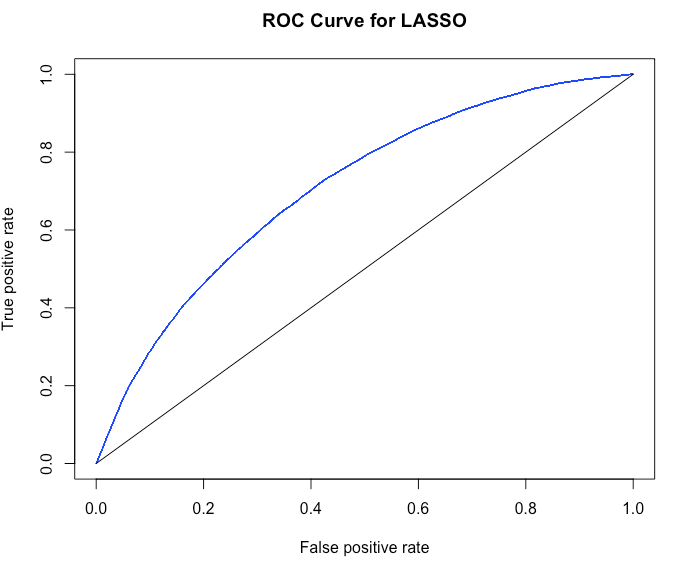
## [1] "Area Under Curve for decision tree is : 0.7083"

**LASSO**

The LASSO (Least Absolute Shrinkage and Selection Operator) is a regression method that involves penalizing the absolute size of the regression coefficients. By penalizing (or equivalently constraining the sum of the absolute values of the estimates) you end up in a situation where some of the parameter estimates may be exactly zero. This is convenient when we want some automatic feature/variable selection, or when dealing with highly correlated predictors, where standard regression will usually have regression coefficients that are 'too large'.

|  |  |
| --- | --- |
| ## Confusion Matrix and Statistics (cut-off = 20%)  ##  ## Actual ## Prediction 0 1 ## 0 29070 3996 ## 1 30233 15771 ##  ## Accuracy : 0.5671  ## Sensitivity : 0.7978  ## Specificity : 0.3428 | Confusion Matrix and Statistics (cut-off = 25%) ##  ## Actual ## Prediction 0 1 ## 0 37565 6578 ## 1 21738 13189 ##  ## Accuracy : 0.6419  ## Sensitivity : 0.6672  ## Specificity : 0.6334 |

|  |  |
| --- | --- |
| Confusion Matrix and Statistics (cut-off = 30%) ##  ## Actual ## Prediction 0 1 ## 0 44505 9294 ## 1 14798 10473 ##  ## Accuracy : 0.6953   Sensitivity : 0.5298  ## Specificity : 0.7505 | Confusion Matrix and Statistics (cut-off = 50%) ##  ## Actual ## Prediction 0 1 ## 0 56951 17157 ## 1 2352 2610 ##  ## Accuracy : 0.7533   Sensitivity : 0.1320  ## Specificity : 0.9560 |



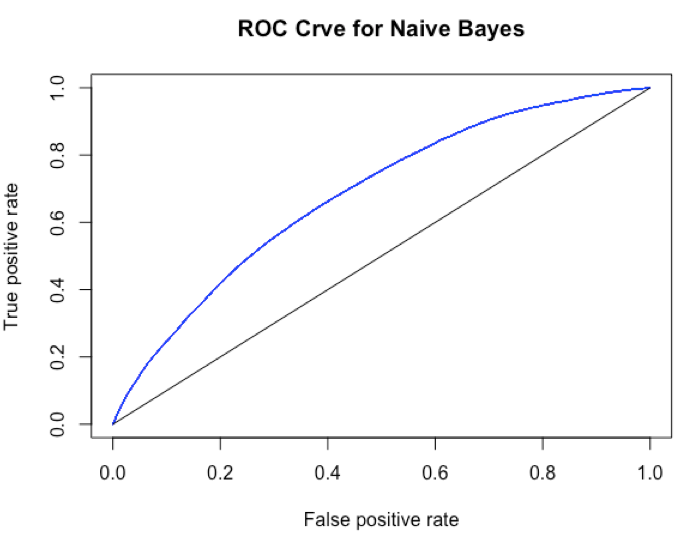
## [1] "Area Under Curve for decision tree is : 0.70818"

**Naïve Bayes**

The Naïve Bayes classifier technique is based on the Bayes theorem and assumes the predicators to be independent, which means knowing the value of one attribute does influence the value of any other attribute. The independence assumption is what makes Naïve Bayes naïve. Naïve Bayes classifiers are easy to build, do not involve any iterative process, and work very well with large datasets. Despite its simplicity, Naïve Bayes is known to have often outperformed other classification methods.

|  |  |
| --- | --- |
| ## Confusion Matrix and Statistics (cut-off = 20%)  ##  ## Actual ## Prediction 0 1 ## 0 31939 5563 ## 1 27364 14204 ##  ## Accuracy : 0.5863  ## Sensitivity : 0.7186  ## Specificity : 0.5286 | Confusion Matrix and Statistics (cut-off = 25%) ##  ## Actual ## Prediction 0 1 ## 0 36429 6950 ## 1 22874 12817 ##  ## Accuracy : 0.6228  ## Sensitivity : 0.6484  ## Specificity : 0.6143 |

|  |  |
| --- | --- |
| Confusion Matrix and Statistics (cut-off = 30%) ##  ## Actual ## Prediction 0 1 ## 0 39773 8137 ## 1 19530 11630 ##  ## Accuracy : 0.6501   Sensitivity : 0.5884  ## Specificity : 0.6707 | Confusion Matrix and Statistics (cut-off = 50%) ##  ## Actual ## Prediction 0 1 ## 0 49202 12484 ## 1 10101 7283 ##  ## Accuracy : 0.7144   Sensitivity : 0.3684  ## Specificity : 0.8296 |



## [1] "Area Under Curve for decision tree is : 0.68321"

# Summarizing Performance

From the performance measures of all the models, we can see that boosting has the largest area under curve of 0.73566. Also in credit risk analysis, accuracy does not play a major role in analysing performance. Predicting a good loan as bad will have less impact on the business than predicting a bad loan as good. In our case sensitivity i.e. the ability of a classifier to predict important class (class = 1) plays a major role. It means how good is our classifier in predicting a default loan. We can see that in most of the cases, ensemble methods outperforms all the other methods. The business implication for this performance summary table is great. A client can see and understand how their model performs at various cut-off values.

Example Scenario: 25% cut-off (loan with default probability more than .25 is termed as bad loan)

In this scenario, looking at the sensitivity, we can say that ensemble method bagging with sensitivity of 0.7493 has the best results followed by random forest with sensitivity of 0.744



# Business Recommendations:

1. Avoid loan with debt-to-income ratio more than 40
2. Loan with greater interest rate gives better return but their probability to go default is even more as the interest rate is highly influenced by the grades provided by Lending Club.
3. Charge more interest rate for borrower with verified source as they are less likely to default.
4. Borrower with Mortgage or rented loan are more likely to default.
5. Better model management that spans the entire modelling life cycle.
6. Data visualization capabilities and business intelligence tools that get important information into the hands of those who need it, when they need it.
7. Implementation of ensemble methods in production.

# Challenges to Successful Credit Risk Management

* **Inefficient data management :** An inability to access the right data when it’s needed causes problematic delays.
* **No group wide risk modelling framework :** Without it, banks can’t generate complex, meaningful risk measures and get a big picture of group wide risk.
* **Constant rework** **:** Analysts can’t change model parameters easily, which results in too much duplication of effort and negatively affects a bank’s efficiency ratio.
* **Insufficient risk tools :** Without a robust risk solution, banks can’t identify portfolio concentrations or re-grade portfolios often enough to effectively manage risk.
* **Cumbersome reporting :** Manual, spreadsheet-based reporting processes overburden analysts and IT.