ALY-6110 Final Project

Lending Club Loan Data Analysis

By

Vastal Vadi

Yash Patel

Nagarjuna Reddy Thiyyagura

Vinod Kumar Chintanippu

Course Name: ALY6110 21303 Data Management & Big Data

Academic term: SEC 03 Winter 2019 CPS Quarter

*Instructor Name:* Valeriy Shevchenko

Northeastern University

**Introduction**

big data analytics not only look at data, but also collect it, organize it, and analyze it for relevant information. The aim is to find patterns, correlations, and information that can help a company make important decisions. Ultimately, the goal of big data analytics is to find all of the knowledge that the data can provide for a company. Scientists, modelers, and many others in the analytics field use big data analytics to sift through large amounts of data that can come from a variety of sources like transactions, web servers, social media, surveys, and even emails.

Peer to Peer is the practice of [lending](https://en.wikipedia.org/wiki/Loan) money to individuals or businesses through online services that match lenders with borrowers. Since peer-to-peer lending companies offering these services generally operate online, they can run with lower [overhead](https://en.wikipedia.org/wiki/Overhead_(business)) and provide the service more cheaply than traditional financial institutions. and most of the peer to peer loans are unsecured personal loans. Steps involved are peer to peer lending platform enable to borrowers to create unsecured personal loans. Investors can search and browse the loan listings on peer to peer lending platform website and select loans they want to invest in based on information supplied about the borrower, amount of loan, loan grade, and loan purpose. Investors make money from the interest and peer to peer lending platform make money on service fee and origination fees.

**Problem Statement**

The lender's investment in the loan is not normally protected by any government guarantee. lenders mitigate the risk of [bad debt](https://en.wikipedia.org/wiki/Bad_debt) by choosing which borrowers to lend to, and mitigate total risk by diversifying their investments among different borrowers based information provided by peer to peer lending platforms. One of the main advantages of person-to-person lending for borrowers can sometimes be better rates than traditional bank rates can offer. The advantages for lenders can be higher returns than obtainable from a savings account or other investments, but subject to risk of loss, unlike a savings account. Interest rates and the methodology for calculating those rates varies among peer-to-peer lending platforms.

But there is risk involving in this process like defaults or bad loans.

In traditional model for predicting the risk involves FICO score based on Payment history, length of credit history, types of credit used and so on and human discretion. With traditional model is not enough for predicting the credit risk.

**Using the Big data analytics for Assessing the risk**

Today the availability of data from number of public websites, public records and personal background information like education status, job title, Home location and other data can provide a more complete picture of a borrower.

By combining both FICO score and additional data, more complete composite of borrower can be analyzed and presented to potential lenders.

With help big data technologies, technical ability and the analytical skills, ability to utilize these newforms of data are very helpful for peer to peer lending platforms to determine the potential risk of a loan and for assigning the credit rating and interest rate.

Big data analytics are useful in looking for patterns, correlations and better assess and price credit risk

In the era of Big Data, it’s these unique processes and analysis that are a catalyst for creating truly disruptive financial services.

With the help of big data, analytical models able to accurately predict credit risk and also provide insights of data for decision making by analyzing the large amount of data.

By analyzing the borrowers based on rating, lower risk borrowers can get loans with less interest rate which is very helpful for borrowers.

**Analysis Method**

For analyzing the large amount loan data for finding the insights, patterns, building the predictive model for assessing deterring the potential risk and analyzing the factors effecting the borrowers credit score, we are using the statistical computing tool R.

R is a [programming language](https://en.wikipedia.org/wiki/Programming_language) and [free](https://en.wikipedia.org/wiki/Free_software) software environment for [statistical computing](https://en.wikipedia.org/wiki/Statistical_computing) and graphics supported by the R Foundation for Statistical Computing. With help rich packages available in we can build model and visualize results effectively.

We have used to below methods for data analysis using R software.

1. Exploratory data Analysis
2. Predictive modelling

We have followed below steps for data analysis using R.

1. Data collection
2. Data Preparation
3. Exploratory Data Analysis
4. Univariate Analysis
5. Multivariate analysis.
6. Predictive Modelling.

**Data collection**

We have collected data from the lending club official website for analysis purpose and data set contains approximately 82K rows and 143 columns of data about borrowers. We used sample of data for analysis.

**Data preparation**

In this project we are analyzing the personal loan payment dataset of Lending Club Corp, LC, available on their website to better understand the best borrower profile for investors.

The dataset covers an extensive amount of information about the borrower like home ownership, job title, previous loan details, FICO score and different other variables.

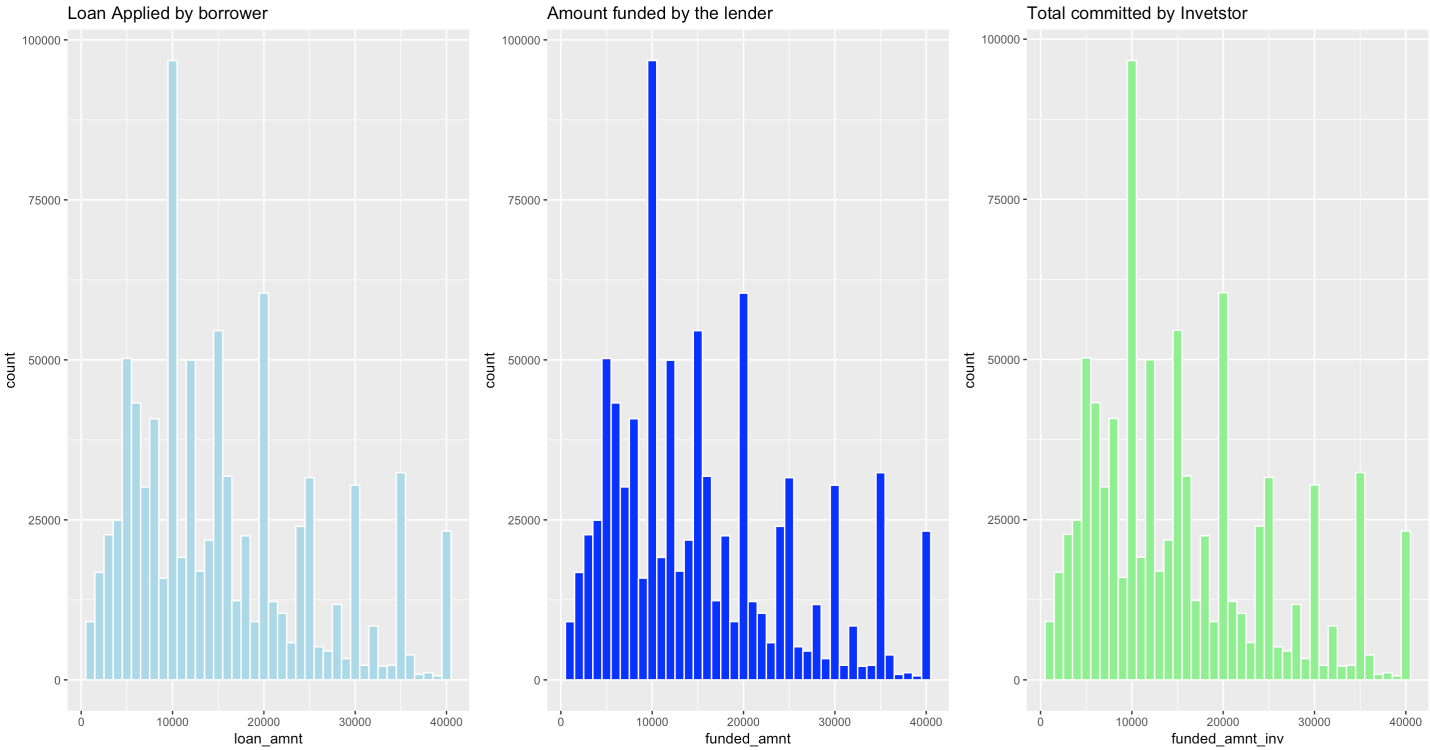
And data set covers historical and quarterly data about the borrowers.

* Exploratory data analysis

In this step we are exploring data in order to understand the insights and underlying pattern in data.

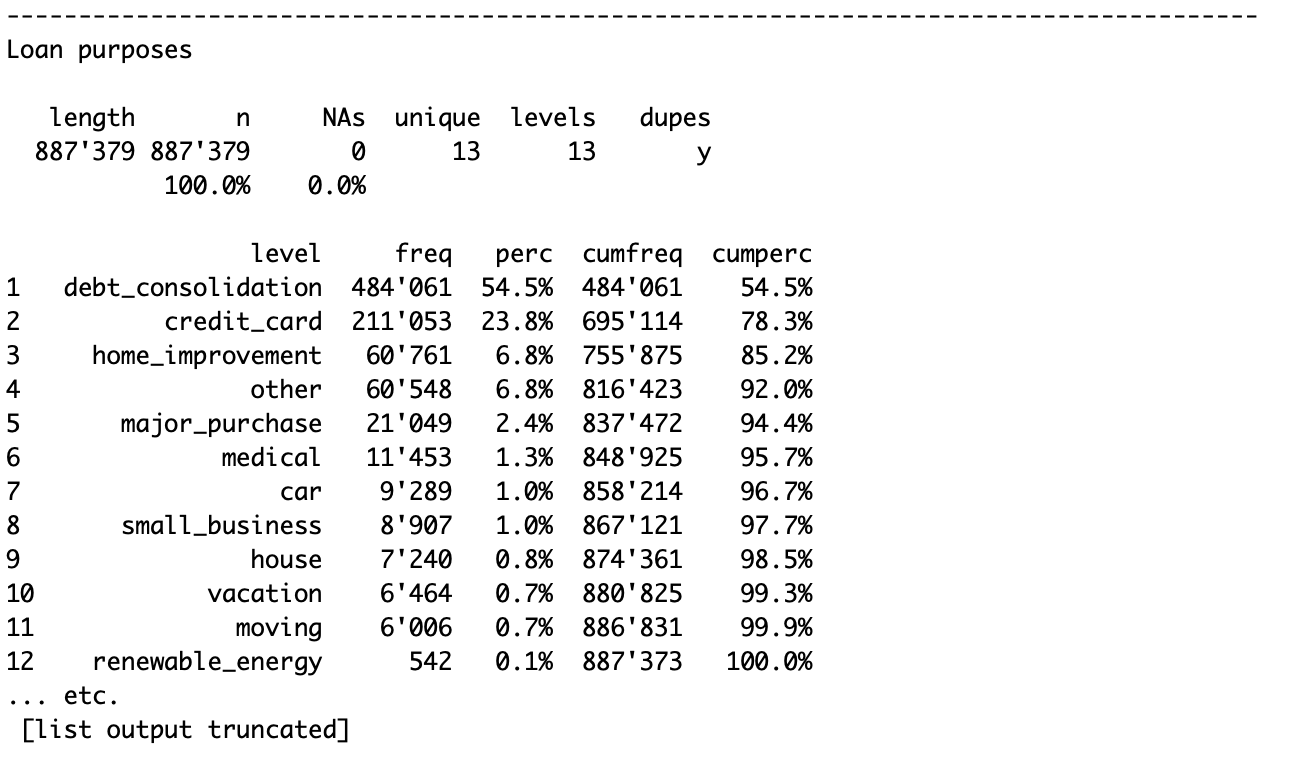
1.Univariate analysis

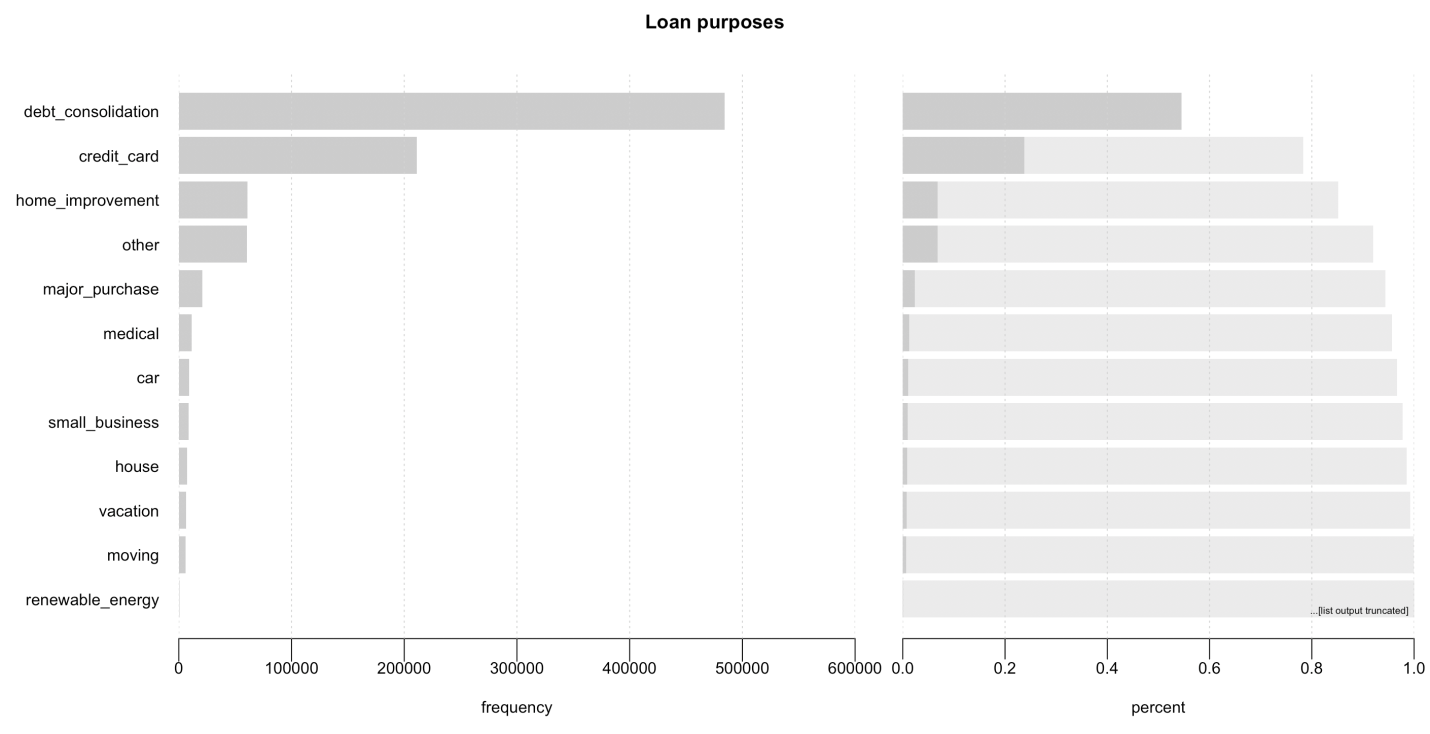
1.1. Understand what amount was mostly issued to borrowers



**Analysis:** From the above graph we can say that most of the loan’s issues in the range of 10,000 t0 20,000 USD.

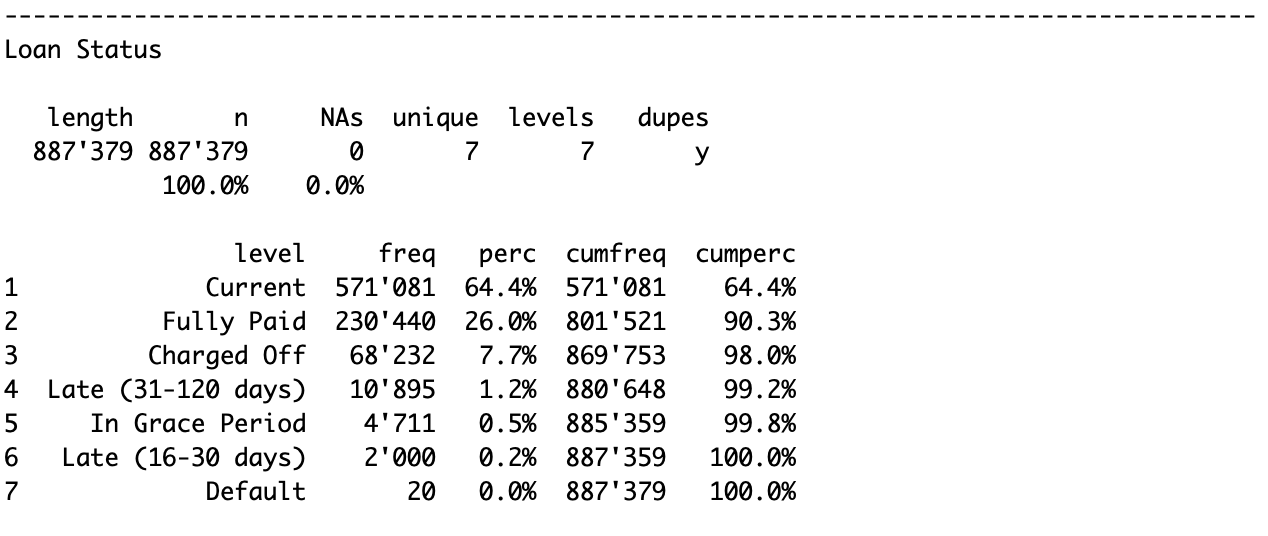
1.2. Loan purpose

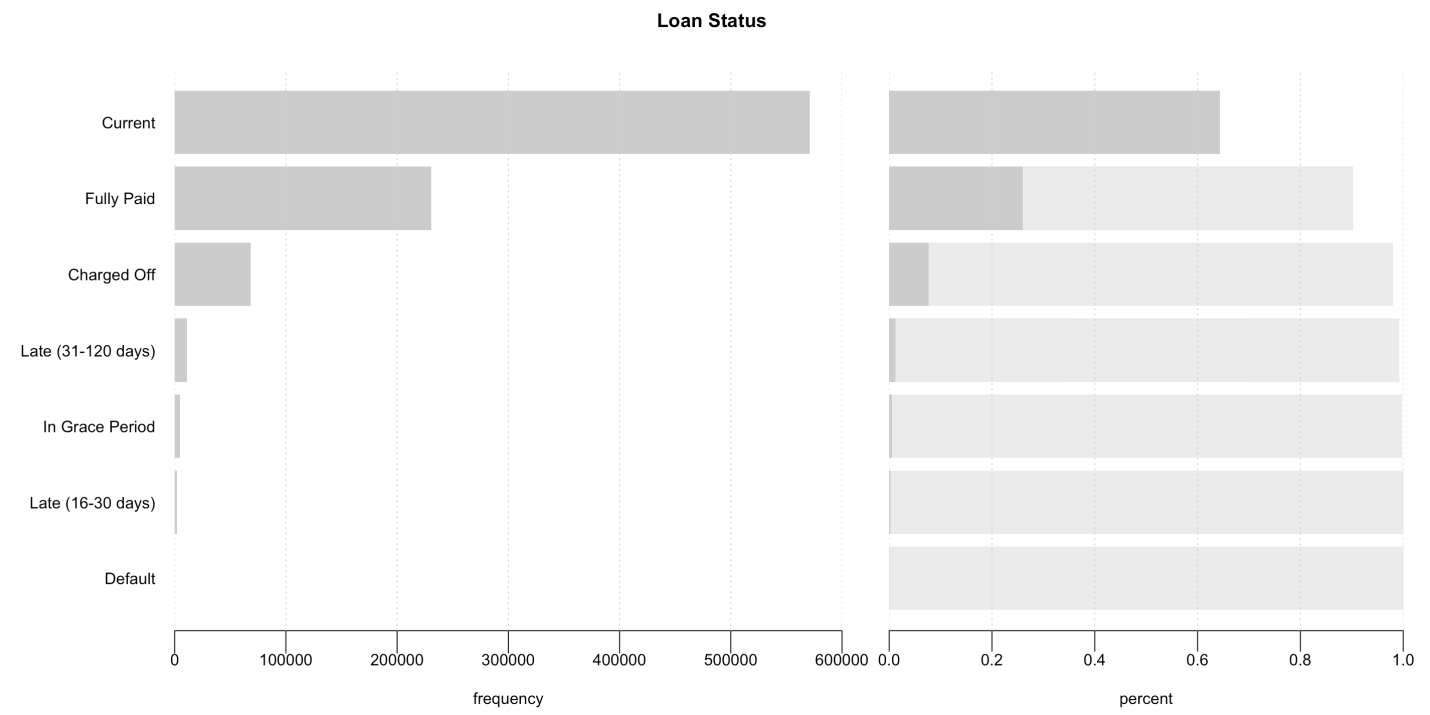




**Analysis:** From the above table we can say that 85% of borrowers taking loans for debt consolidation, credit card and home improvement.

1.3 Loan Statutes



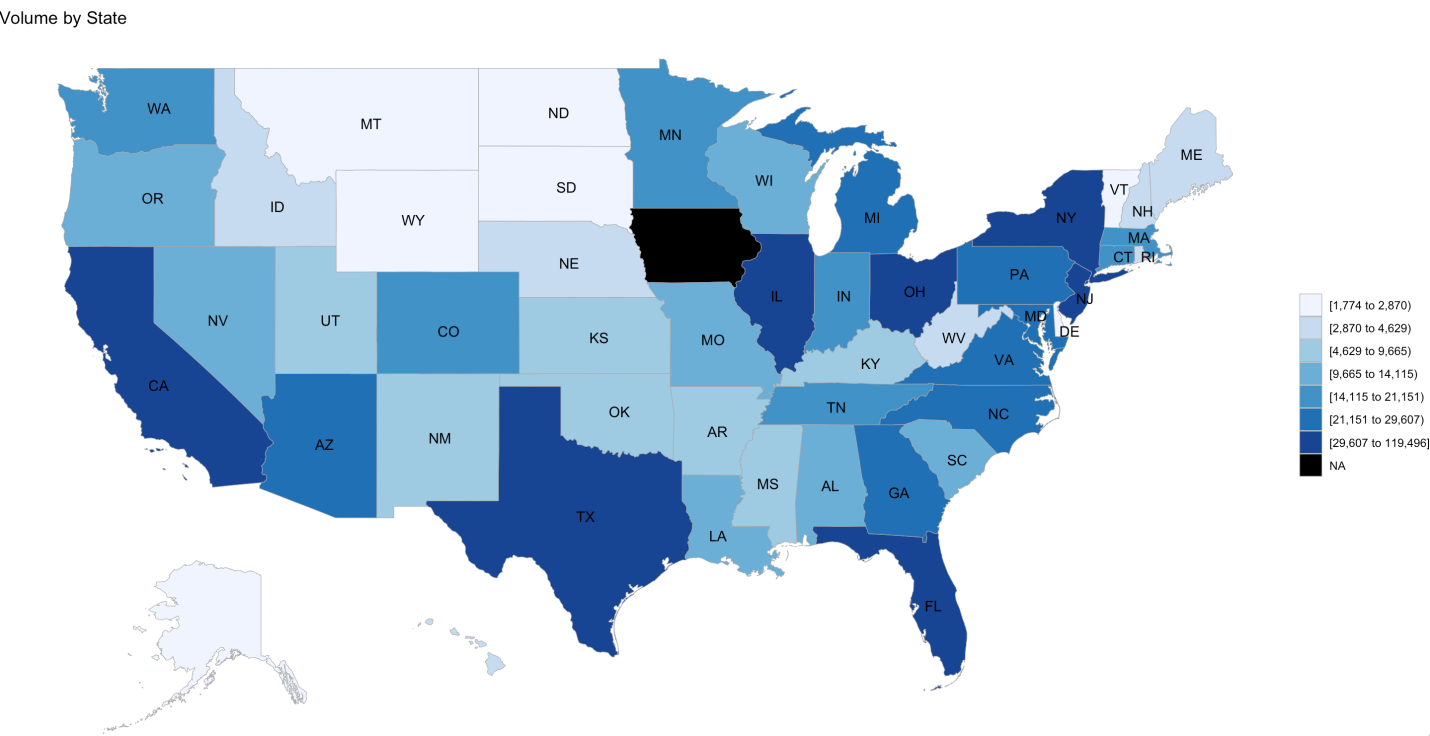


**Analysis:** Loan status like current, full paid and are good loans and loan status like charged off, late, grace period and default are bad loans.

From the above table, we can say

So, we can say that 90.4 % loans are not defaulted and only 9.6 % loans are bad loans.

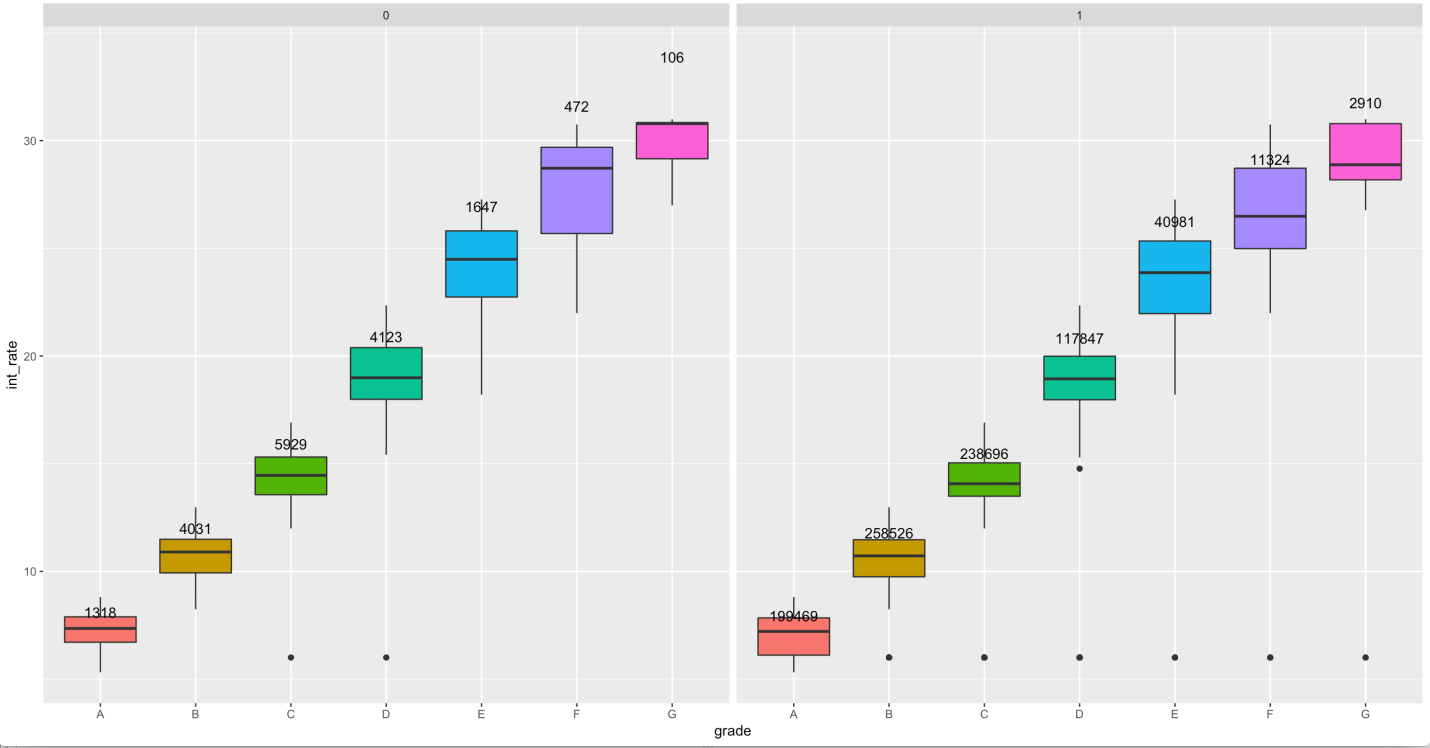
1.4 volume by state



**Analysis:** From the above map, we can say that most of the borrowers are from the CA, TX, FL, IL OH, NY states.

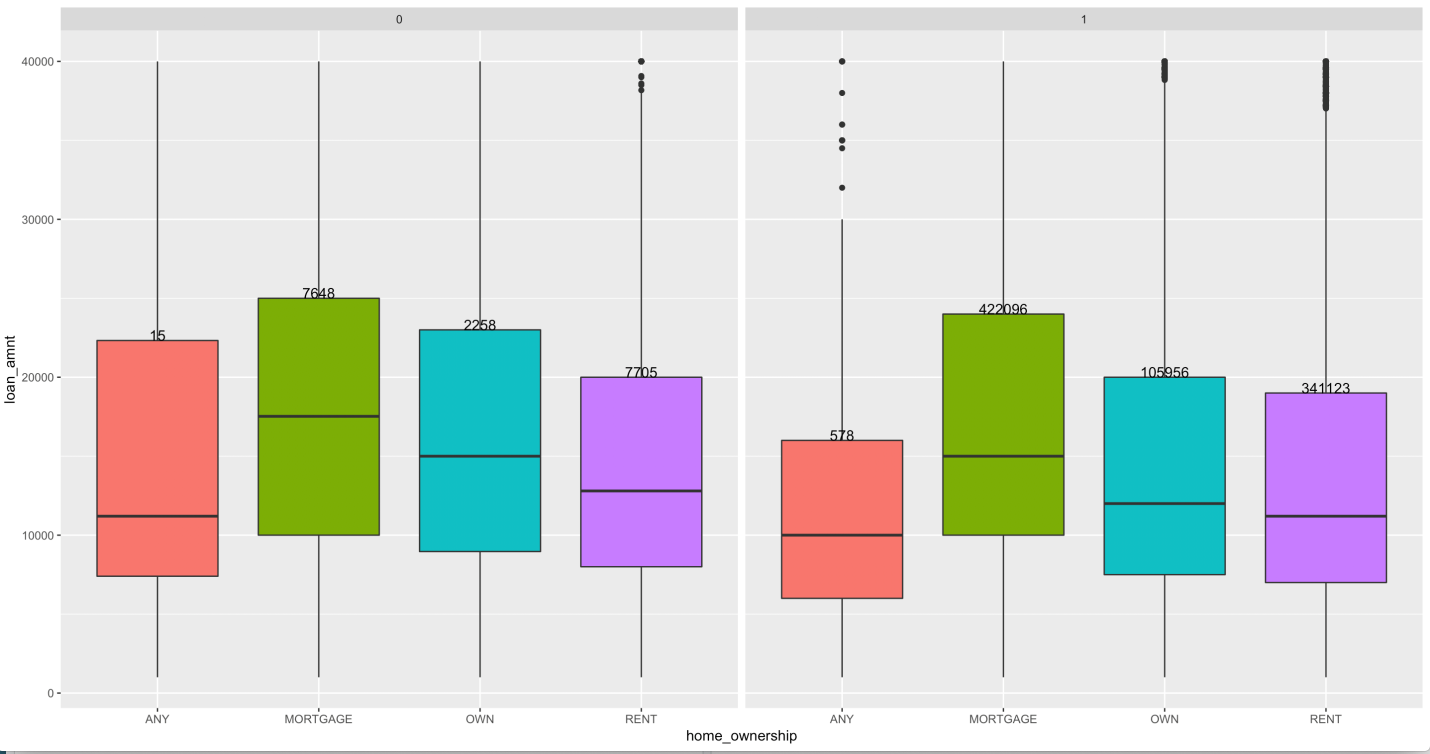
2. Multivariate analysis.

2.1. Grade Vs Interest Rate



**Analysis:** From the Above graph (0 indicates bad loan, 1 indicates good loan) we can say that loan quality increases, interest rate decreases and no of loans belongs to good to medium quality loans. A grade being the highest quality loan and G grade is lowest.

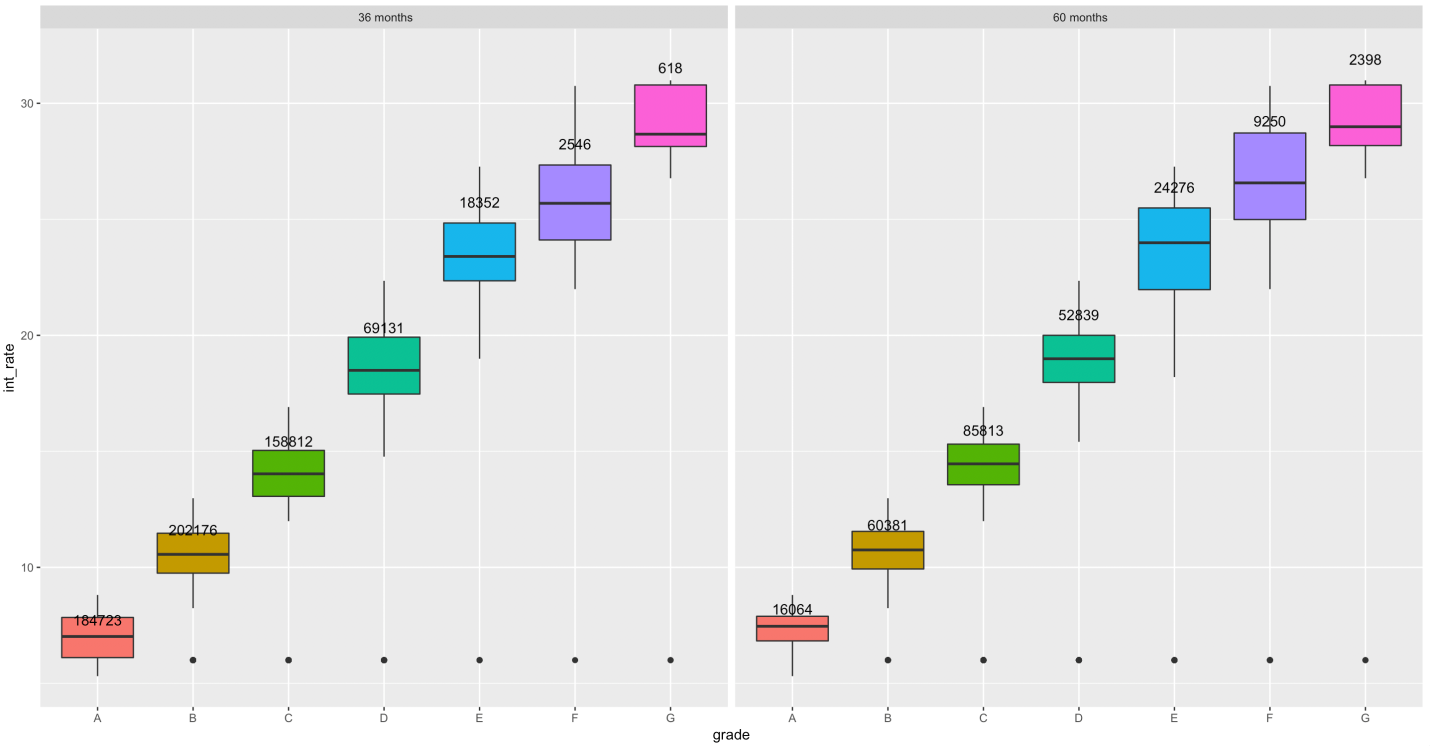
2.2. Homeownership Vs Loan amount



**Analysis:**

From the above Graph,

* Borrowers who are on rented or on mortgaged homes, took most of the loans.
* Most defaulters are on rented homes.
  1. Grade Vs Interest Rate Vs Tenure.



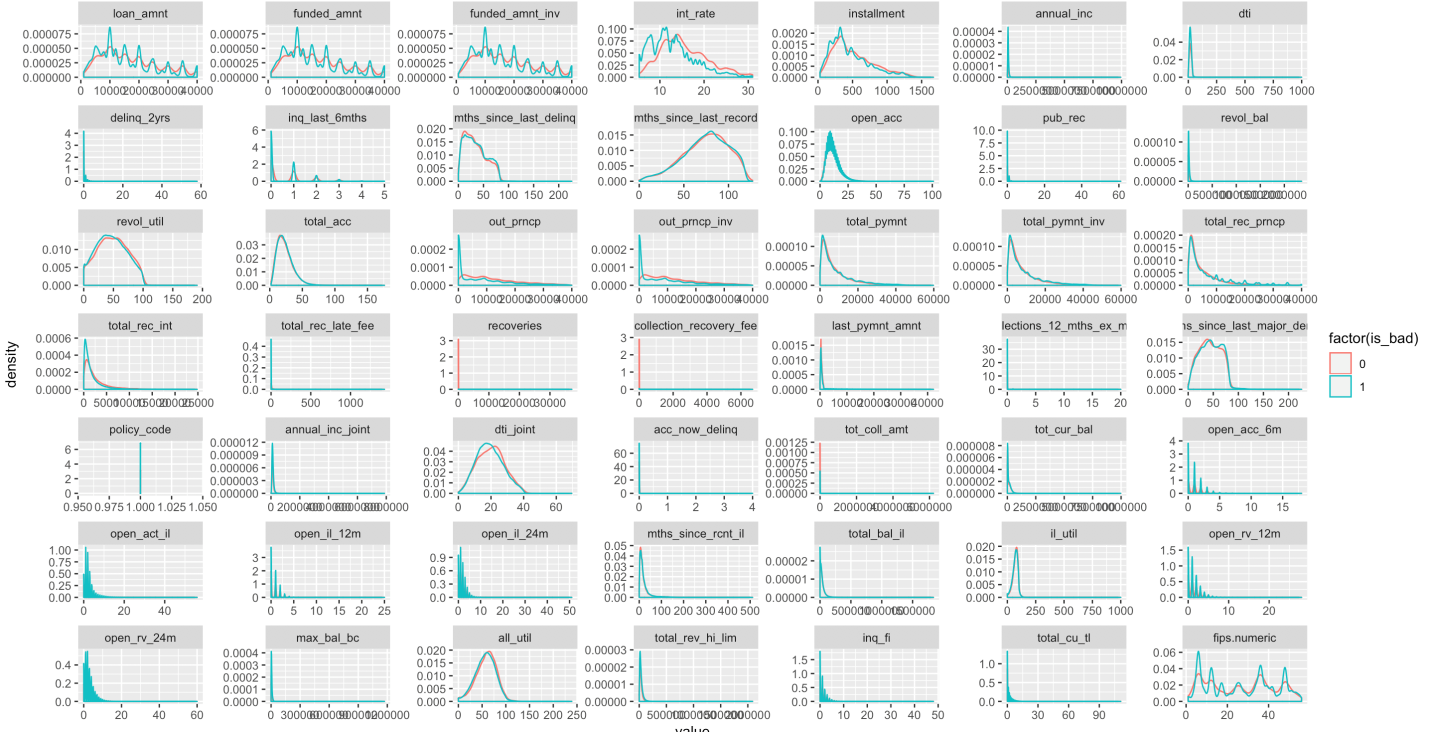
**Analysis:**

From The above graph,

* Major loans are taken for 36 months tenure
* More good quality belongs to 36 months.
* Medium to low quality loans are taken for 60 months tenure.

1. **Predictive modelling**
   1. Selecting the predictors

Data set is huge and many of the variables will not be helping to predict the defaulted loans, we are choosing few variables based on above analysis.



**Analysis:**

We can use above graph can be used to for choosing the variables for predictive models.

We used below variables to predict the whether or not loan is defaulted.

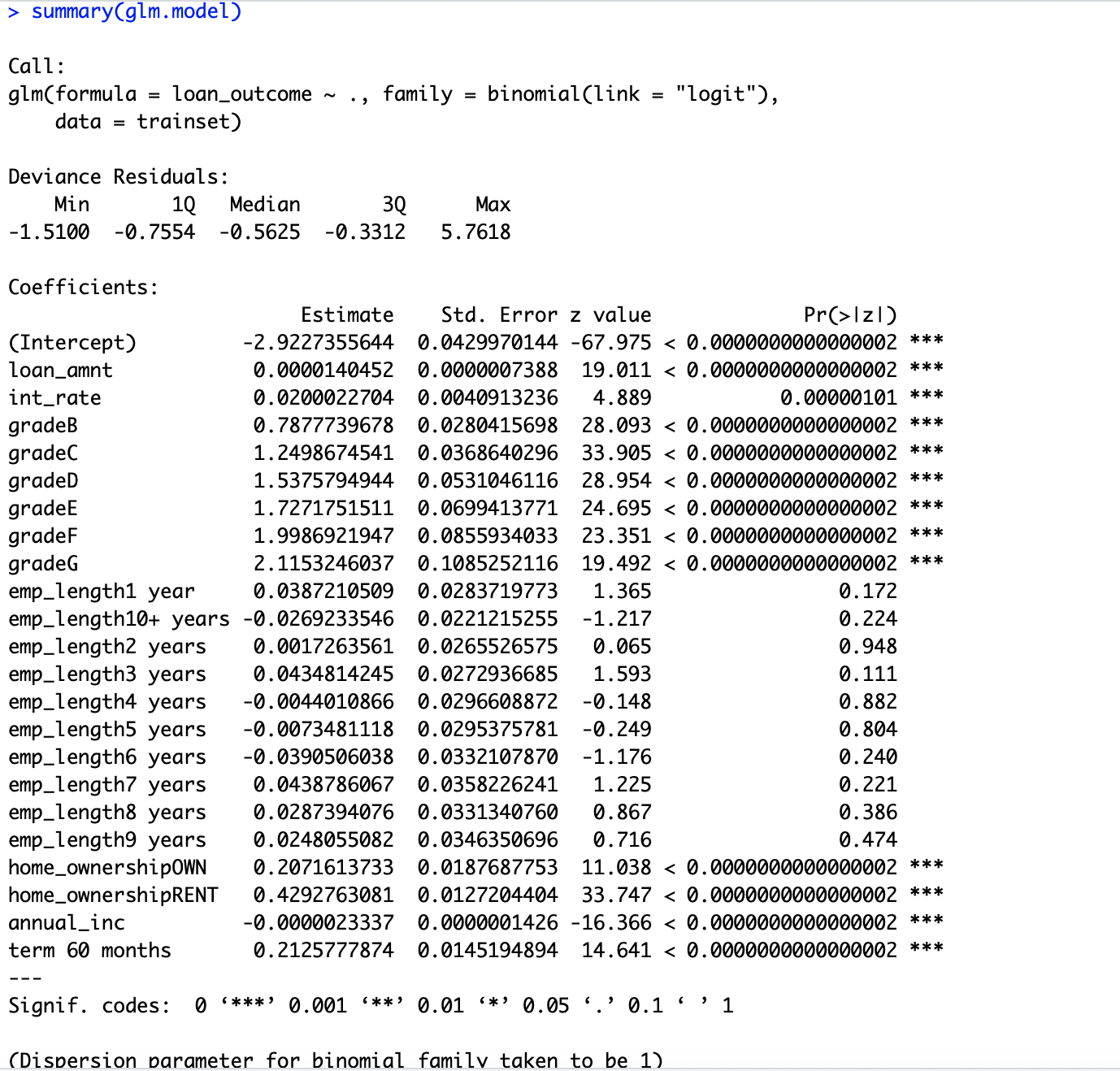
loan\_status, loan\_amnt, int\_rate, grade, emp\_length, home\_ownership, annual\_inc, term.

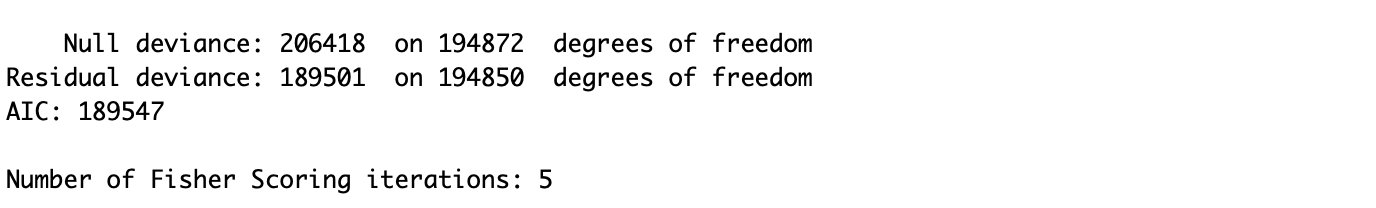
* 1. Partitioning data sets

We need to partition data sets into Training and validation sets based on 70% and 30% ratio. Training set will be used to train the models and validation set is used to validate the accuracy of the model.

* 1. Applying logistic regression model.

Using the regression model to predict the target variable.

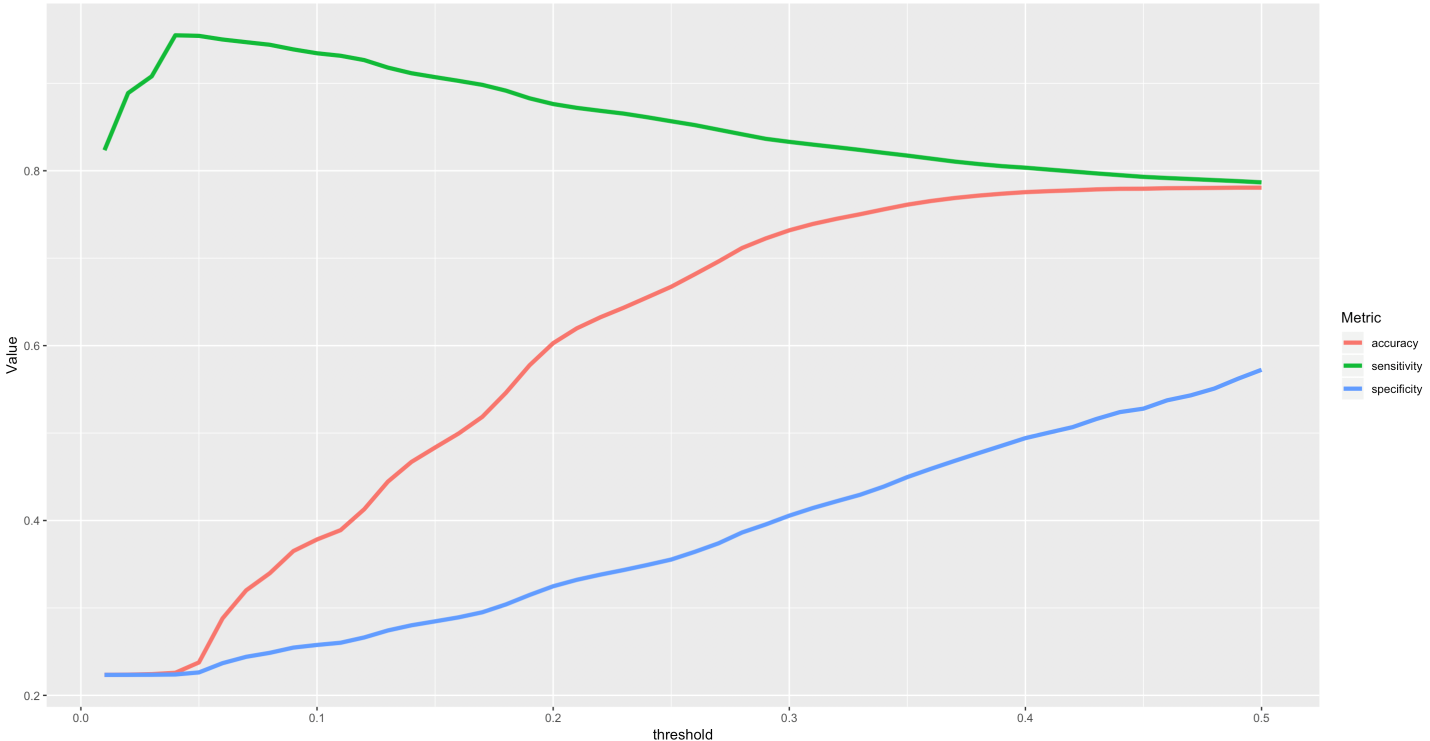




* 1. Interpretation of coefficients
* Predictors which have p-values less than 0.05 have significant effect on target variable.
* Variables which have positive coefficients means the probability of defaulting on the given credit varies directly with these variables and varies inversely proportional with variables which have negative coefficients.
  1. Predicting the target variable

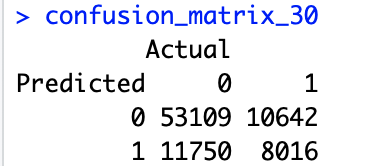
Based on above model we need to predict the target variable by using test data set.

* 1. Finding the threshold of the model.



Based on value of accuracy setting the threshold value to 0.3.

* 1. Finding the accuracy of the model.



From the above confusion matrix table, Accuracy of the model is 0.73189

* 1. Validation of prediction results.

Area under the curve: 0.7037

Different threshold was used to decide If the loan should be granted or not. Threshold values 0f 30% gave a good accuracy of 73.18 %. The area under the curve also gives a measure of accuracy which is 70%.

**Conclusion**

Lending Club’s data is a great source of information on personal credit. Additionally, this data set is bound to grow exponentially over the next years. Analyzing the borrower’s data will help to analyze the credit risk. It will help both borrowers and investors. From the analysis we found that amount most of the loan’s issues, Reasons for borrowing, Percentages of bad loans will help to understand about the borrowers and by Predictive modelling we found which factors clearly explaining whether loan is bad one or good one. So that investors can invest in low risk owners and low risk owners will get loans for low interest rates with the help predictive modelling.

**References**

Loan Payment dataset. Retrieved from <https://www.lendingclub.com/info/download-data.action>.

https://www.houseofbots.com/news-detail/4104-1-how-automation-is-changing-data-science-and-machine-learning-pattern

https://www.btelligent.com/en/portfolio/big-data/

**# Source Code**

require(dplyr)

require(ggplot2)

require(ggpubr)

require(gridExtra)

require(choroplethrMaps)

require(DT)

library(gridExtra)

library(dplyr)

library(ggpubr)

library(choroplethrMaps)

library(plyr)

library(DT)

library(ggthemes)

library(pROC)

#Disabling scientific notation in R

options(scipen = 999)

# raeding data from the file

data1<-read.csv(file.choose(),header=T,nrows = 887379)[ ,3:73]

dim(data1)

head(data1)

View(data1)

# merge the loan book with the state names

data1 <- merge(data1, state.regions, by.x = "addr\_state", by.y = "abb")

colnames(data1)

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*Explorarory data Analysis \*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

## 1.univariate data analysis

# Loan amount applied borrower

P1<-ggplot(data1, aes(loan\_amnt,fill=I("lightblue"), col=I("white"))) + geom\_histogram(binwidth=1000) +ggtitle("Loan Applied by borrower")

# total amount commited to loan at that point of time.

P2<-ggplot(data1, aes(funded\_amnt,fill=I("blue"), col=I("white"))) +geom\_histogram(binwidth=1000)+ ggtitle("Amount funded by the lender")

#amount given by investors

P3<-ggplot(data1, aes(funded\_amnt\_inv,fill=I("lightgreen"), col=I("white"))) +geom\_histogram(binwidth=1000) + ggtitle("Total committed by Invetstor")

grid.arrange(P1, P2,P3, nrow = 1)

## Loan status.

#checking the differen loan status

distinct(data1, loan\_status)

Desc(data1$loan\_status, main = "Loan Status", plotit = TRUE)

# Loan Purpose

Desc(data1$purpose, main = "Loan purposes", plotit = TRUE)

# volume by State

install.packages("choroplethrMaps")

library(choroplethrMaps)

library(choroplethr)

data(state.regions)

head(state.regions)

state\_by\_volume <-

data1 %>% group\_by(region) %>%

summarise(value = n())

state\_choropleth(state\_by\_volume, title = "Volume by State")

# Multivarriate analysis

## Grade vs Interst rate

ggplot(data1,aes(grade,int\_rate,fill=grade))+geom\_boxplot(show.legend = F)+

facet\_grid(.~is\_bad)+

stat\_summary(geom = "text",fun.data=function(x) {return(c(y=median(x)\*1.1,label=length(x)))})

##home ownership vs laon amount

ggplot(data1,aes(home\_ownership,loan\_amnt,fill=home\_ownership))+geom\_boxplot(show.legend = F)+facet\_grid(.~is\_bad)+

stat\_summary(geom = "text",fun.data=function(x) {return(c(y=quantile(x,.75)[[1]],label=length(x)))},aes(vjust="bottom"))

## Grade Vs. Interest Rate Vs Tenure

ggplot(data1,aes(grade,int\_rate,fill=grade))+geom\_boxplot(show.legend = F)+facet\_grid(.~term)+

stat\_summary(geom = "text",fun.data=function(x) {return(c(y=median(x)\*1.1,label=length(x)))})

# for choosing the variables for Predictive modelling

# 'bad' statuses

bad\_loanindicators <- c("Charged Off ","Default",

"In Grace Period",

"Default Receiver",

"Late (16-30 days)",

"Late (31-120 days)")

# assign certain statuses to a 'bad' ('0') group

data1$is\_bad <- ifelse(data1$loan\_status %in% bad\_loanindicators, 0,

ifelse(data1$loan\_status=="", NA, 1))

View(data1$is\_bad)

# figure out which columns are numeric so that we can look at the distribution

numeric\_cols <- sapply(data1, is.numeric)

library(reshape2)

data1.lng <- melt(data1[,numeric\_cols], id="is\_bad")

p <- ggplot(aes(x = value, group = is\_bad, colour = factor(is\_bad)),

data = data1.lng)

p + geom\_density() +

facet\_wrap(~variable, scales="free")

library(DT)

data1 %>%

filter(is\_bad == '0') %>%

select(loan\_amnt, loan\_status) %>%

datatable(., options = list(pageLength = 10))

#\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*\*

# Predictive Modeling

# Selecting the predictors and preparing data

data2<- select(data1,loan\_status, loan\_amnt, int\_rate, grade, emp\_length, home\_ownership,annual\_inc, term)

View(data2)

dim(data2)

sapply(data2 , function(x) sum(is.na(x)))

str(data2)

data2<-filter(data2,!is.na(annual\_inc) ,

!(home\_ownership %in% c('NONE' , 'ANY')) ,

emp\_length != 'n/a')

data2 = data2 %>% mutate(loan\_outcome = ifelse(loan\_status %in% c('Charged Off' , 'Default') ,

1,

ifelse(loan\_status == 'Fully Paid' , 0 , 'No info')))

barplot(table(data2$loan\_outcome) , col = 'lightblue')

loan2<-data2 %>%select(-loan\_status) %>%filter(loan\_outcome %in% c(0 , 1))

head(loan2)

# Partiotioning data Set

loan2$loan\_outcome = as.numeric(loan2$loan\_outcome)

idx = sample(dim(loan2)[1] , 0.70\*dim(loan2)[1] , replace = F)

trainset = loan2[idx , ]

testset = loan2[-idx , ]

# Fit logistic regression

glm.model <-glm(loan\_outcome ~ . , trainset , family = binomial(link = 'logit'))

summary(glm.model)

glm.model

# Prediction on test set

preds<-predict(glm.model , testset , type = 'response')

preds

# Density of probabilities

ggplot(data.frame(preds) , aes(preds)) +

geom\_density(fill = 'lightblue' , alpha = 0.4) +

labs(x = 'Predicted Probabilities on test set')

k = 0

accuracy = c()

sensitivity = c()

specificity = c()

for(i in seq(from = 0.01 , to = 0.5 , by = 0.01)){

k = k + 1

preds\_binomial = ifelse(preds > i , 1 , 0)

confmat = table(testset$loan\_outcome , preds\_binomial)

accuracy[k] = sum(diag(confmat)) / sum(confmat)

sensitivity[k] = confmat[1 , 1] / sum(confmat[ , 1])

specificity[k] = confmat[2 , 2] / sum(confmat[ , 2])

}

threshold<-seq(from = 0.01 , to = 0.5 , by = 0.01)

data<-data.frame(threshold , accuracy , sensitivity , specificity)

head(data)

library(tidyr)

# Gather accuracy , sensitivity and specificity in one column

ggplot(gather(data , key = 'Metric' , value = 'Value' , 2:4) ,

aes(x = threshold , y = Value , color = Metric)) +

geom\_line(size = 1.5)

#confusion Matrix

preds.for.30<-ifelse(preds > 0.3 , 1 , 0)

confusion\_matrix\_30 = table(Predicted = preds.for.30 , Actual = testset$loan\_outcome)

confusion\_matrix\_30

# Validation of predicted results

library(pROC)

# Area Under Curve

auc(roc(testset$loan\_outcome , preds))

# Plot ROC curve

plot.roc(testset$loan\_outcome , preds , main = "Confidence interval of a threshold" , percent = TRUE ,

ci = TRUE , of = "thresholds" , thresholds = "best" , print.thres = "best" , col = 'blue')