Data607 Final Project

Description:

A Bad Debt or Loan is a monetary amount owed to a creditor that is unlikely to be paid and, or which the creditor is not willing to take action to collect because of various reasons, often due to the debtor not having the money to pay. Iam trying to build a bad loan model that can be used by the investors to easily decide whether to finance the borrower for new loans. I will be using Machine Learning Random Forest or some other classification algorithms.

Data Description:

Loan data of Lending Club(https://www.lendingclub.com/info/download-data.action) from 2007-2011. I will import the .csv file and use it to build model.

I will also use unemployment rate data from Bureau of Labor Statistics(https://data.bls.gov/map/MapToolServlet).I may do web scrapping or load the data to MongoDB.

Data preparation:

The documentation published in the “Data Dictionary “on the Lending Club website was very helpful in understanding and knowing the variables and their description.

The unemployment data was taken from Bureau of Labor Statistics(https://data.bls.gov/map/MapToolServlet)

Data preparation code snippet:

Writing to MongoDB:

Writing the unemployment data to the MongoDB

```
c=mongo(collection="unemp", db="upemp")
```

```
c$drop()
c$insert(unemp)
```

```
## List of 5
##  $ nInserted  : num 1
##  $ nMatched   : num 0
##  $ nRemoved   : num 0
##  $ nUpserted  : num 0
##  $ writeErrors: list()
```

```
alldata <- c$find('{}')
alldata
```

```
##    addr_state  AL  AK  AZ  AR   CA  CO  CT  DE  DC  FL  GA  HI  ID  IL  IN
## 1 un_emp_rate 8.3 7.4 8.8 7.8 11.2 8.2 8.3 7.2 9.9 9.1 9.8 6.8 7.8 9.4 8.8
##    IA  KS  KY  LA  ME MD  MA  MI  MN  MS  MO  MT  NE   NV  NH  NJ  NM  NY
## 1 5.3 6.1 8.6 7.5 7.7  7 6.8 9.4 5.9 9.4 7.6 6.4 4.2 12.3 5.4 9.2 7.6 8.6
##    NC  ND OH  OK  OR  PA   RI SC  SD  TN  TX UT  VT  VA  WA  WV  WI  WY
## 1 9.8 3.3  8 5.6 9.2 7.8 11.2 10 4.4 8.1 7.2  6 5.2 6.4 8.8 7.5 7.3 5.6
```

```
alldata1 <- gather(alldata, "addr_state", "un_emp_rate")
head(alldata1)
```

```
##   addr_state un_emp_rate
## 1         AL         8.3
## 2         AK         7.4
## 3         AZ         8.8
## 4         AR         7.8
## 5         CA        11.2
## 6         CO         8.2
```

Merge the loans data and unemployment data:

```
loans_data <- merge(loans, alldata1, by="addr_state", all.x=TRUE)
count(loans_data)
```

```
## # A tibble: 1 x 1
##       n
```

```
##   <int>
## 1 38652
```

Cleaning the data with mostly na values:

There were lots of columns with na(invalid) values. Cleaning the data with na values.

Code snippet:

```
#remove fields that are mostly NA
pc <- sapply(loans_data, function(x) {
  t1 <- 1 - sum(is.na(x)) / length(x)
  t1 < .8
})
df <- loans_data[,pc==FALSE]
head(df)
```

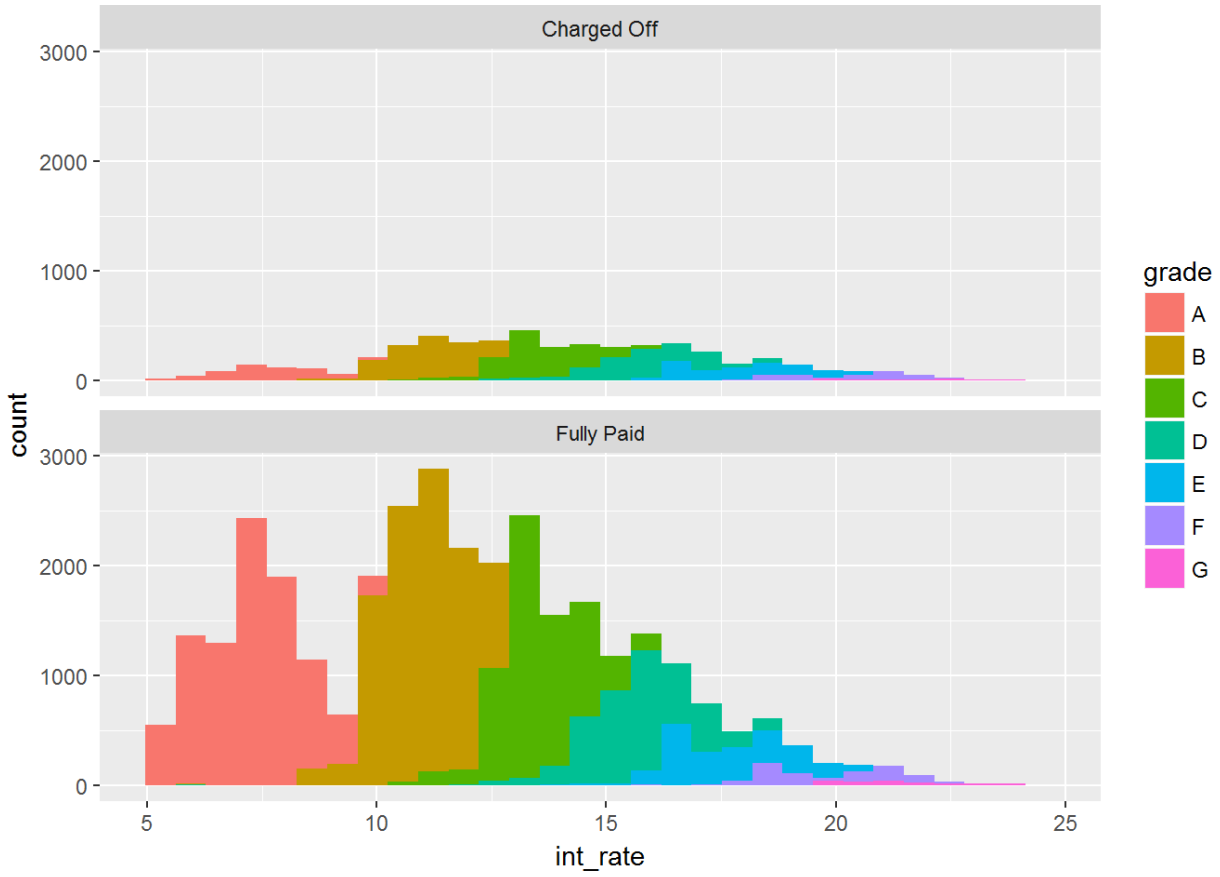
Next I did the analysis of loan status:

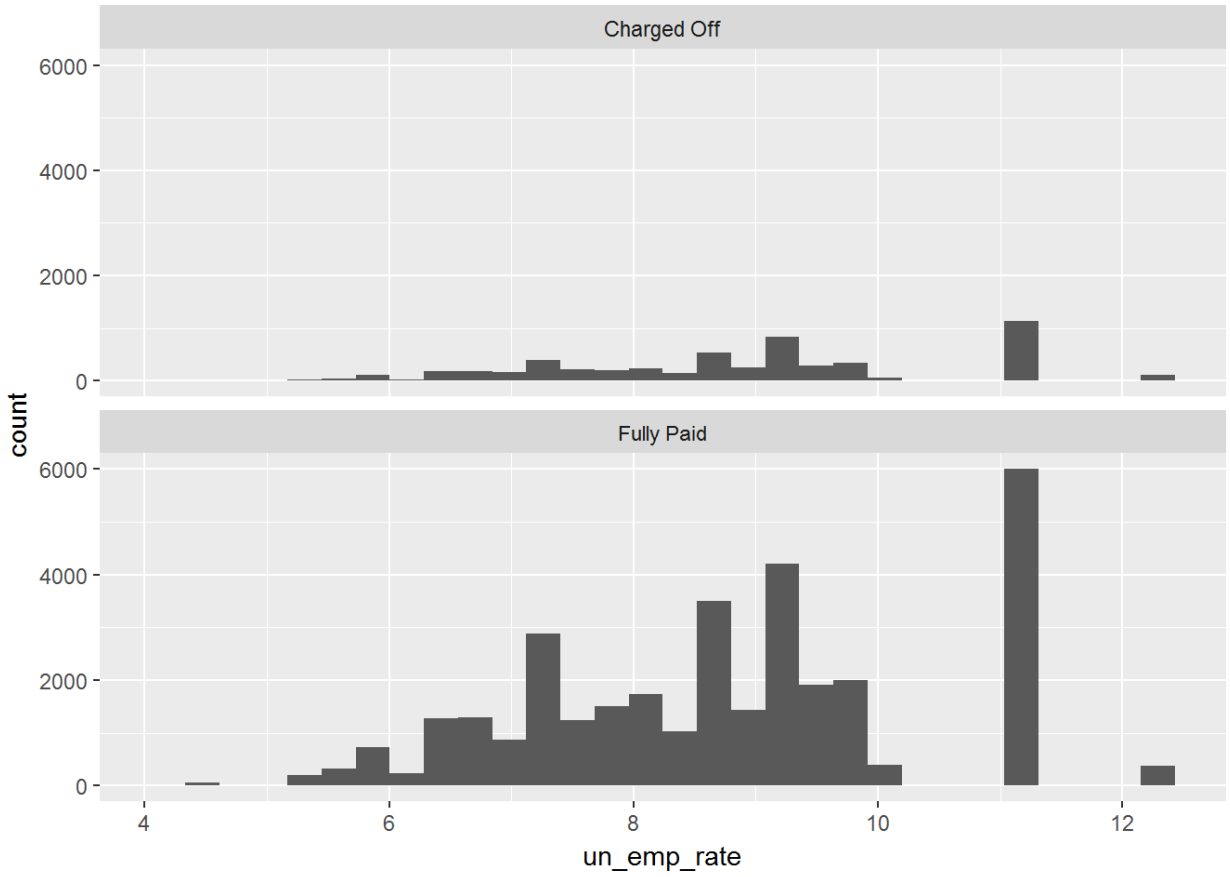
I was able to segregate the loans into two categories : “Charged Off” and “Fully Paid”

This is where we have to build the model to predict.

Exploratory Analysis:

Checking loan status against the loan grades. Most of the lower grade loans failed to pay back.


Charged Off
Fully Paid
count
3000
2000
1000
0
5
10
15
20
25
int_rate
grade
A
B
C
D
E
F
G



Finding out the numeric columns:

Code snippet:

```
numeric_cols <- sapply(df, is.numeric)
df.lng <- melt(df[,numeric_cols], id ="bad_loans")
head(df.lng)
```

We need to find all the numeric data that we can utilize to decide on the predictors, so iam doing numeric analysis. Ploting the graph for clear understanding.

```
p <- ggplot(aes(x=value, group=bad_loans, colour=factor(bad_loans)), data=df.
lng)
```

```
p + geom_density() +
  facet_wrap(~variable, scales="free")
```

Removing outliers:

We need to remove the outliers so that it should not impact by making wrong predictions.

Code snippet:

```
inc_outliers <- which(df$annual_inc > 1000000)
df <- df[-inc_outliers,]
```

Building the Models

```
#Sampling the data
set.seed(123)


sample <- runif(nrow(df)) > 0.70
train <- df[sample==FALSE,]
test <- df[sample==TRUE,]


table(train$bad_loans)
```

We first built the logistics regression Model:

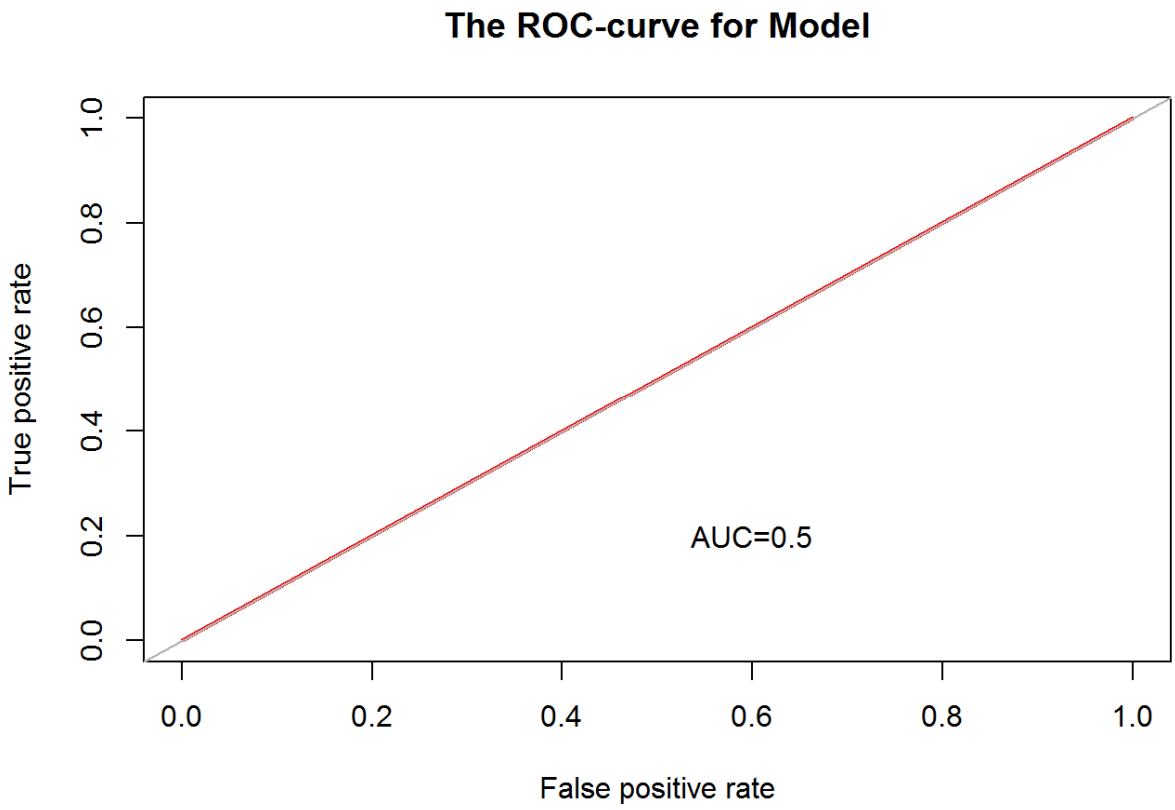
```
#Building the logistic regression model


logistic_regressor <- glm(bad_loans ~ loan_amnt + int_rate + installment + an
nual_inc + dti +
```

```
revol_bal + revol_util + total_acc + un_emp_rate, family = "binomial", data =
train)

summary(logistic_regressor)
```

Seeing the ROC below I decide to resample. The AUC was just 0.5.



I did the resampling for balancing the sample between good and bad loans:

```
#Improving the sampling by balancing between good and bad

improved_train <- ROSE(bad_loans ~ loan_amnt + int_rate + installment + annua
l_inc + dti +

revol_bal + revol_util + total_acc + un_emp_rate, data = train, seed = 1)$dat
a

table(improved_train$bad_loans)
```

Then I built new logistic model:

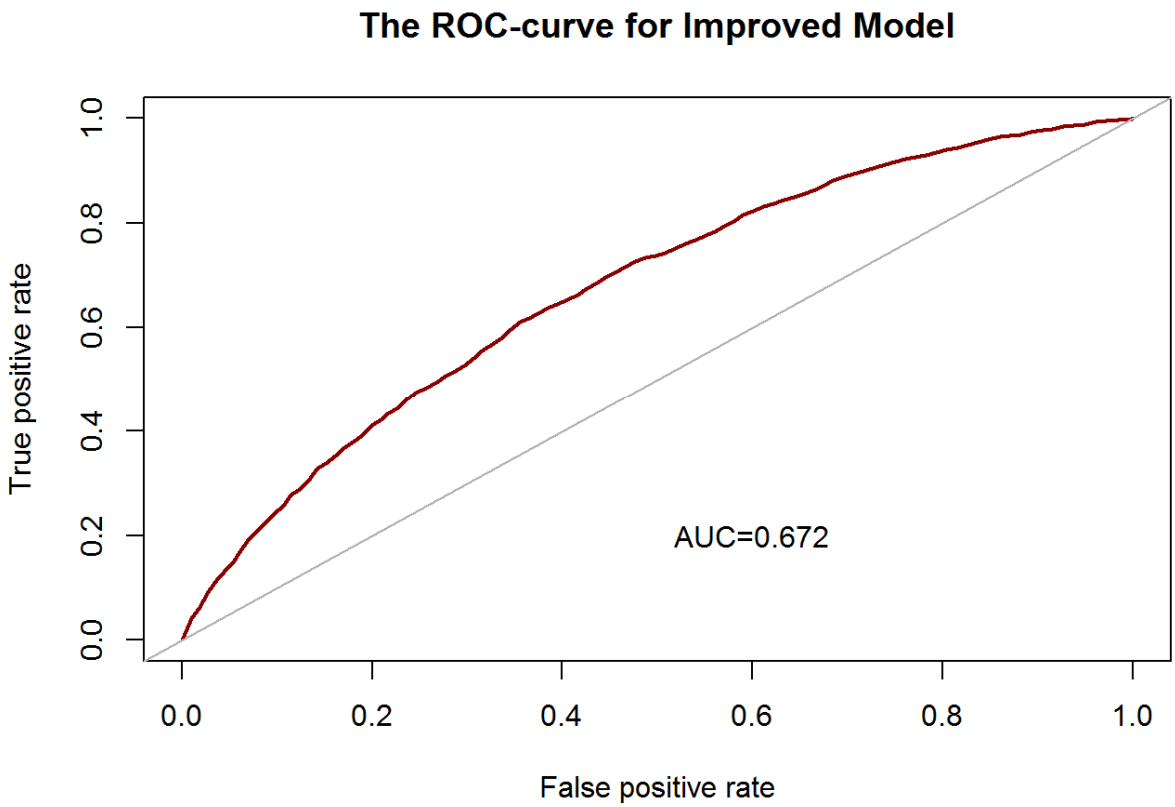
```
#Building new logistic regression model

improved_regressor <- glm(bad_loans ~ loan_amnt + int_rate + installment + an
nual_inc + dti +

revol_bal + revol_util + total_acc + un_emp_rate, family = "binomial", data =
improved_train)

summary(improved_regressor)
```

ROC for new model:



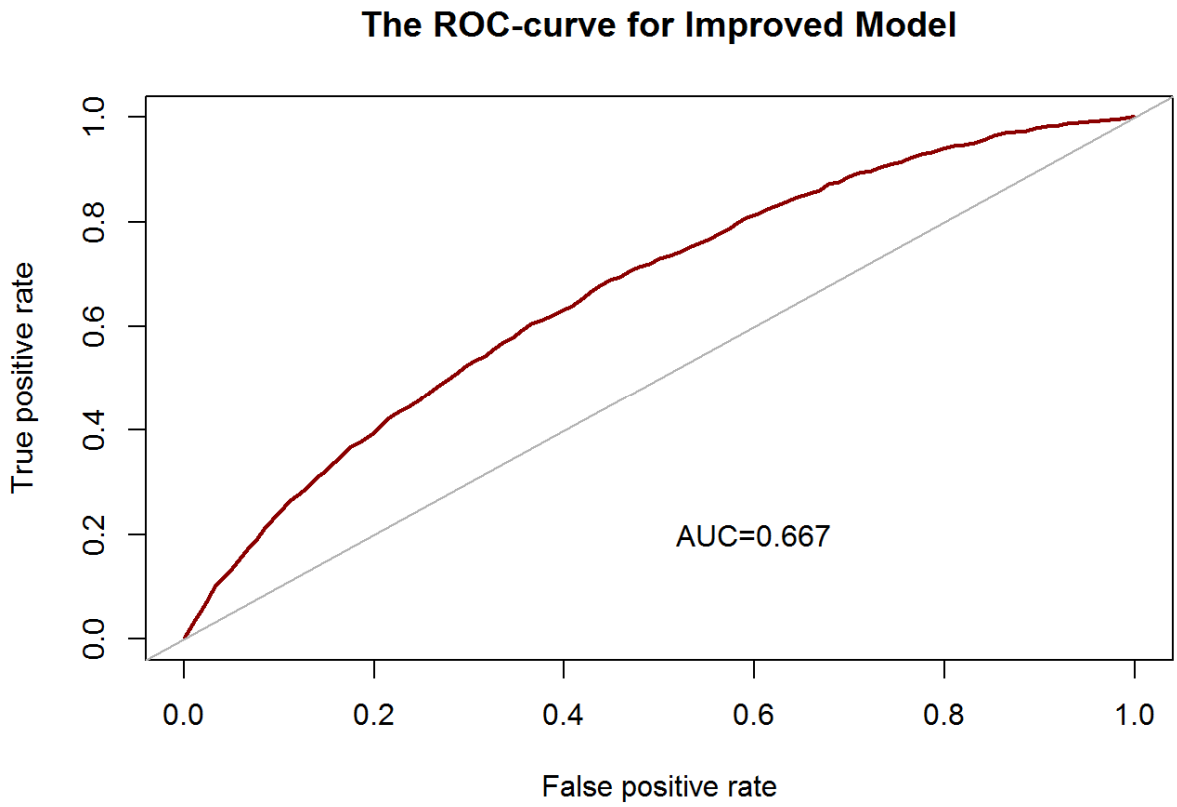
I also tried Random Forest Model:

```
#Using random forest
```

```
rf <- randomForest(bad_loans ~ loan_amnt + int_rate + installment + annual_in
c + dti  + revol_bal + revol_util + total_acc

+ un_emp_rate, type="classification", data=improved_train, importance=TRUE, n
a.action=na.omit)
```

ROC for Random Forest Model:



Conclusions:

I have developed a model using logistic regression and Random Forest to predict if a borrower will repay the loan based on historical data provided by Lending Club and to help investors when deciding which investment strategy to choose.

I think both Logistic and Random Forest Model have similar outcomes. I would say the data is very imbalanced, so very difficult to predict with high accuracy.