Analysis of Bank Loan Status

Viveksinh Solanki
Ronald Fernandes
Yatri Kalathia

Course: FE-582
Sem: Spring 2020
Stevens Institute of Technology

1. Introduction

Financial institutions like banks incur significant losses due to default of loans. This has led to tightening up of loan underwriting and increased loan rejection rates. The need for the better credit risk scoring model is also raised by there institutions. This warrants a study to estimate the determinants of loan default. So, it is very important for any bank to check the background of its customer before lending a loan. Based on customer's background, bank takes the decision whether to lend a loan or not to that customer. Doing so will ensure that clients capable of repayment are not rejected and important determinants can be identified which can be used for minimizing the default rates.

We are interested in finding out the probable parameters which might lead to loan defaults. Hence, we have chosen to work on this problem. We aim to identify the probable parameters by applying various data analysis methods and machine learning models. Our final model will be able to predict whether the customer will default a loan or not based on given parameters.

2. Research Question(s)

- Whether the customer will default on a loan given their financial details?
- Which are the factors that might lead to a loan default?
- Which customers are not capable of repayment of loans?

3. Dataset

Source: Bank Loan status - Kaggle

This dataset is obtained from Kaggle. There are total 100,000 records with 19 columns(features) having numerical as well as categorical features. We are trying to classify **Loan Status.**

Loan ID	String denoting the ID of loan
Customer ID	String denoting the ID of customer
Loan Status	Boolean with following values: Charged Off - Defaulted Fully Paid - Paid
Current Loan Amount	Amount of the Loan
Term	Boolean with following values: 1) Short Term 2) Long Term

Credit Score	Credit score of the customer
Annual Income	Annual Income of the customer
Years in current job	Customer's total number of years in current job
Home Ownership	Specifies the ownership of home in which customer is living: 1) Have Mortgage 2) Home Mortgage 3) Own 4) Rent
Purpose	Purpose of the loan
Monthly Debt	Monthly debt of the customer
Years of Credit History	# of years of customer's credit history
Months since last delinquent	# of months since last delinquent
Number of Open Accounts	# of open accounts of customer
Number of Credit Problems	# of credit problems of customer
Current Credit Balance	Customer's current credit balance
Maximum Open Credit	Customer's maximum preapproved credit
Bankruptcies	# of bankruptcies customers went through till now
Tax Liens	# of tax liens of customer till now

4. EDA

```
rm(list=ls())

setwd("C:/Users/Ron/Desktop/Stevens/FE582_financial_data_science/Project/my_dataset")

train <- read.csv('credit_train.csv', na.strings = "")

train[train=='NA'] <- NA

#head(train)

summary(train[3])

##       Loan.Status
##  Charged Off:22639
##  Fully Paid :77361
##  NA's       :  514

summary(train[4])
```

```
##  Current.Loan.Amount
##  Min.   :   10802
##  1st Qu.:  179652
##  Median :  312246
##  Mean   :11760447
##  3rd Qu.:  524942
##  Max.   :99999999
##  NA's   :514
```

```
summary(train[5])
```

```
##          Term
##  Long Term :27792
##  Short Term:72208
##  NA's      :  514
```

```
summary(train[6])
```

```
##   Credit.Score
##  Min.   : 585
##  1st Qu.: 705
##  Median : 724
##  Mean   :1076
##  3rd Qu.: 741
##  Max.   :7510
##  NA's   :19668
```

```
summary(train[7])
```

```
##  Annual.Income
##  Min.   :    76627
##  1st Qu.:   848844
##  Median :  1174162
##  Mean   :  1378277
##  3rd Qu.:  1650663
##  Max.   :165557393
##  NA's   :19668
```

```
summary(train[8])
```

```
##  Years.in.current.job
##  10+ years:31121
##  2 years  : 9134
##  3 years  : 8169
##  < 1 year : 8164
##  5 years  : 6787
##  (Other)  :36625
##  NA's     :  514
```

```
summary(train[9])
```

```
##        Home.Ownership
##  HaveMortgage :  214
##  Home Mortgage:48410
##  Own Home     : 9182
##  Rent         :42194
##  NA's         :  514
```

```
summary(train[11])
```

```
##   Monthly.Debt
##  Min.   :     0
##  1st Qu.: 10214
##  Median : 16220
##  Mean   : 18472
##  3rd Qu.: 24012
##  Max.   :435843
##  NA's   :514
```

```
summary(train[18])
```

```
##   Bankruptcies
##  0      :88774
##  1      :10475
##  2      :  417
##  3      :   93
##  4      :   27
##  (Other):   10
##  NA's   :  718
```

```
# More than 50000 rows i.e more than 50% of data in months deliquent column i
s null
train <- train[,-13]

train <- train[!is.na(train$Loan.ID),]

#Around 18000 duplicate loan IDs.
nrow(train[duplicated(train$Loan.ID),])
```

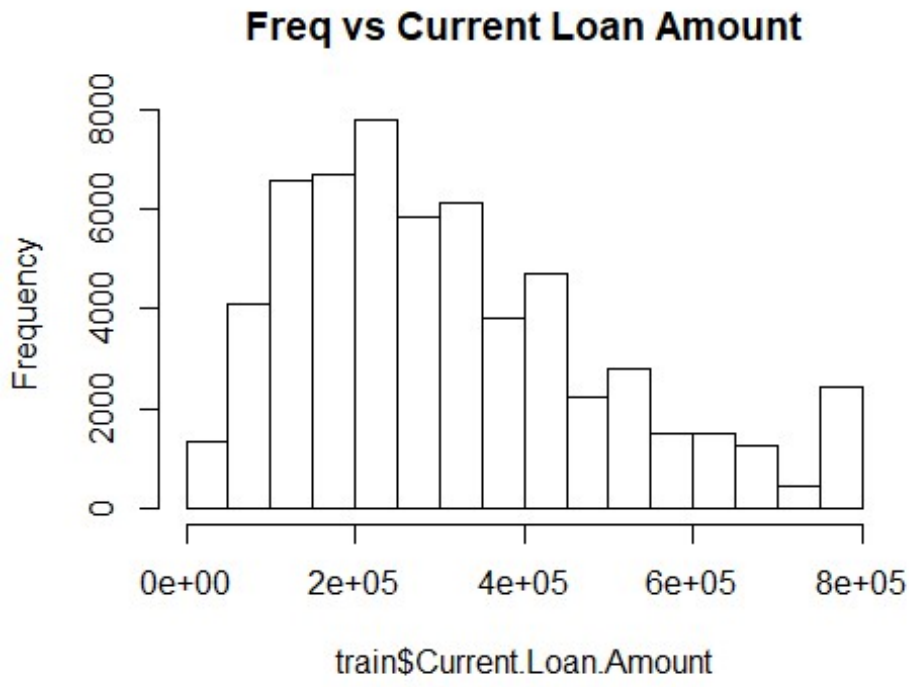
```
## [1] 18001
```

```
# So, only 82000 records are unique.
length(unique(train$Loan.ID))
```

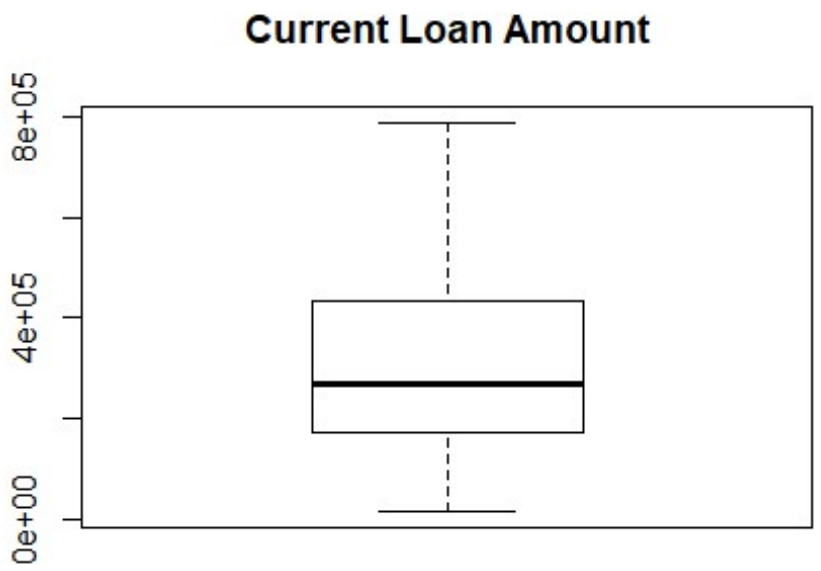
```
## [1] 81999
```

```
train <- train[complete.cases(train),]

#Removing loan amount outlier
train <- train[train$Current.Loan.Amount != 99999999,]

#Since credit score and annual income are the most important factors,
#Hence when both are nulls at the same time it should be removed
```

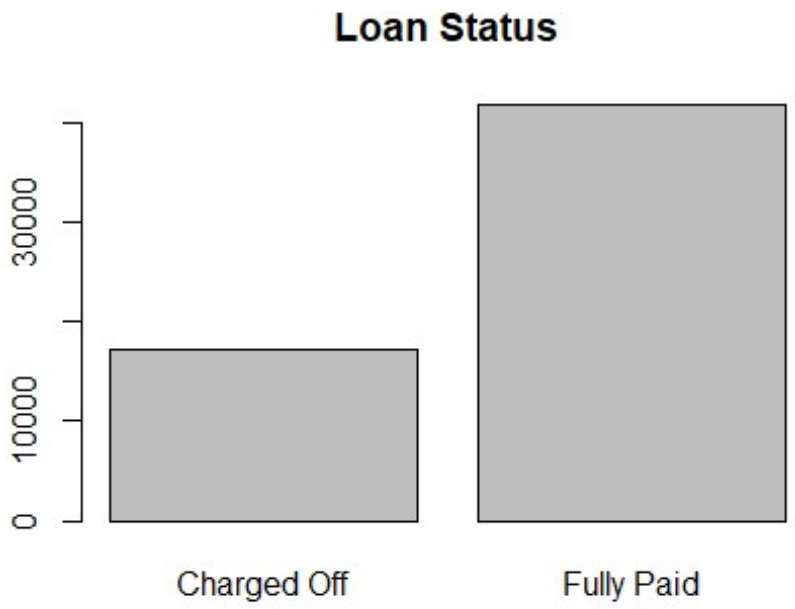
```r
train <- train[!is.na(train$Credit.Score) & !is.na(train$Annual.Income),]

#Removing full duplicated rows
train <- train[!duplicated(train),]

hist(train$Current.Loan.Amount, main="Freq vs Current Loan Amount")
```



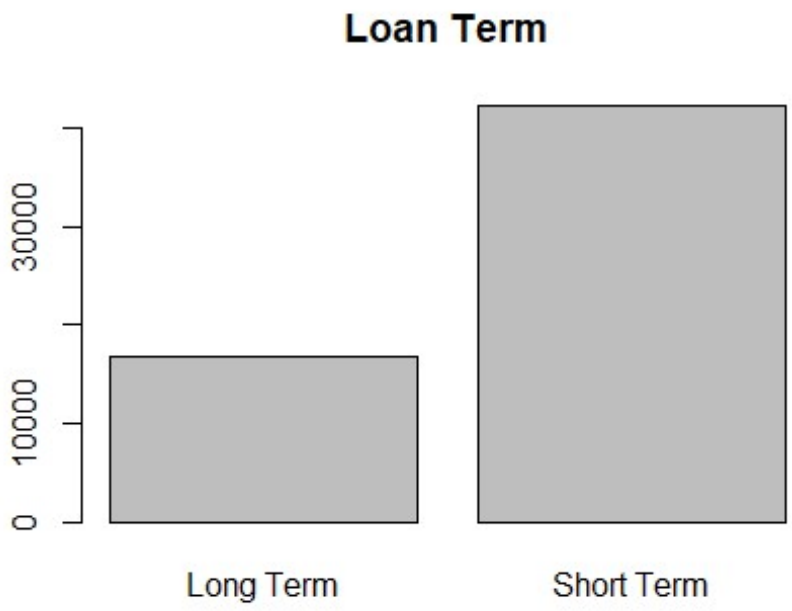
```r
boxplot(train$Current.Loan.Amount, main="Current Loan Amount")
```



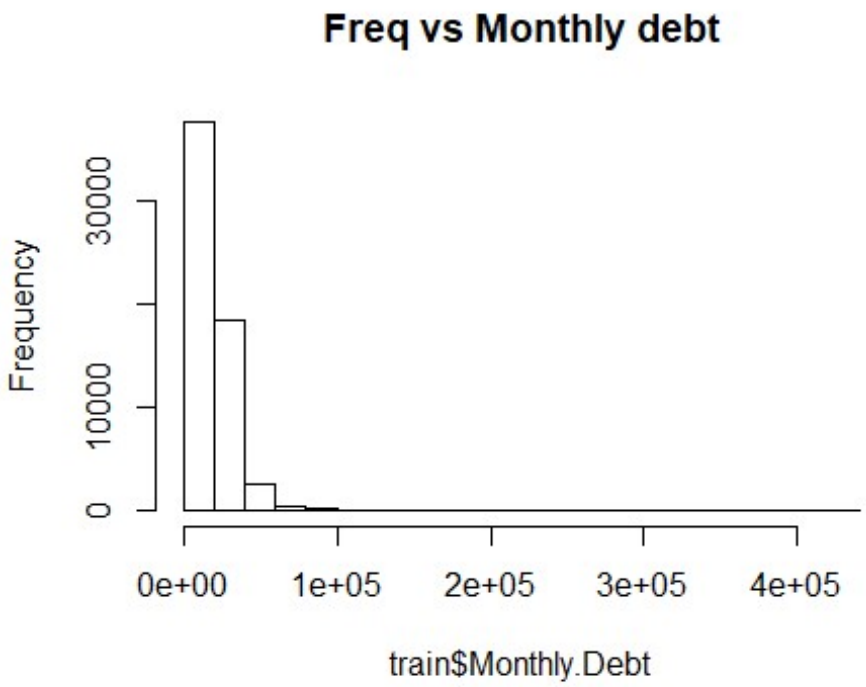

```r
barplot(table(train$Loan.Status), main = "Loan Status")
```



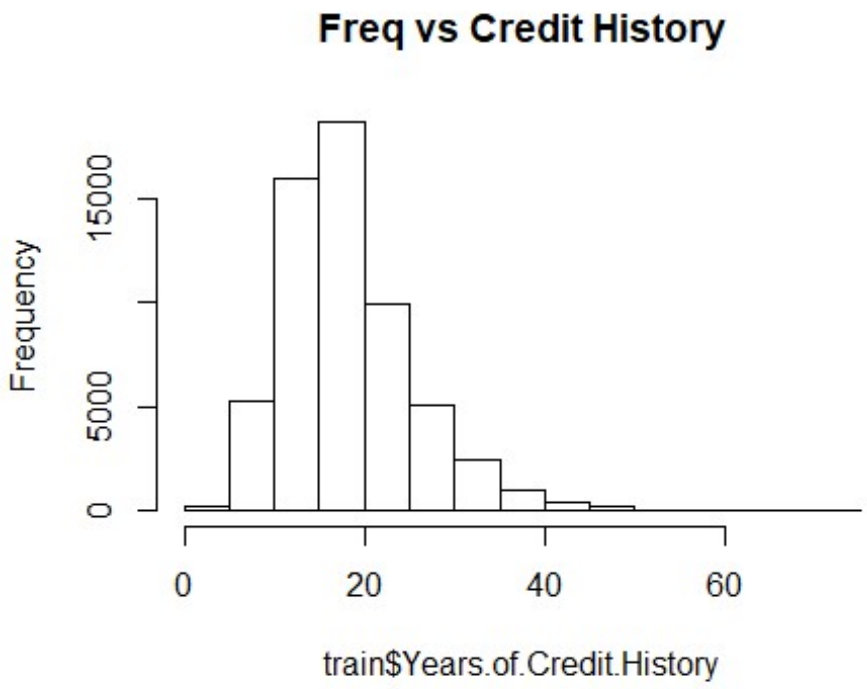
```r
barplot(table(train$Term), main = "Loan Term")
```



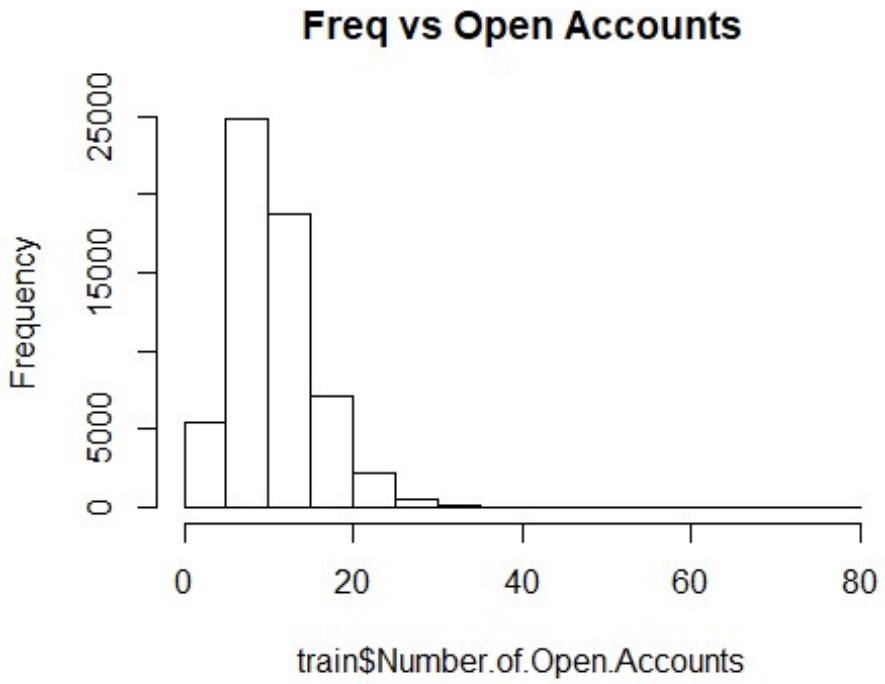

```r
hist(train$Monthly.Debt, main = "Freq vs Monthly debt")
```



```r
hist(train$Years.of.Credit.History, main = "Freq vs Credit History")
```




```r
hist(train$Number.of.Open.Accounts, main = "Freq vs Open Accounts")
```



5. Methods

I. Handling Missing Values
II. Outlier Detection
III. Data Visualization
IV. Classification (Logistic Regression, kNN, Random Forest)
V. Clustering (K-Means)

6. References

- Dataset link: https://www.kaggle.com/zaurbegiev/my-dataset