Predicting Loan Defaults with Logistic Regression

Isuri Willaddara Gamage

July 30, 2020

## Executive Summary

Modeling credit risk for both personal and business loans is of major importance for banks and financial institutions. The likelihood that a borrower will default is a key component in getting to a measure for credit risk. This study shows the application of logistic regression to develop functional model to predict the risk of default in loans and achieve maximum profitability by determining the necessary risk of defaulted loans over the potential for profit.

First, we explore and transform data to use for predictive model and then build and optimize the models. The data was split into two separate sets in which to both build and validate the model and get the accuracy scores for the best selected model. The proposed model has been analyzed it in terms of correct prediction percent of fully paid and default loans status.

The determined model provides and overall accuracy of 67% and by implementing this model, the bank can increase their potential profit by over $2M or 523%. However, the trade off of the profit comes with the price of denying some of the loans that would have been fully paid, the loan’s status that was incorrectly predicted is 33% (model’s accuracy of fully paid loan prediction is 67%).

The model we have built has proven to have a capability of predicting the loan fully paid type of loan based on the information collected from applicant. By putting this model into action, the bank can reduce the default loans and significantly increase the total profit. Even though the suggested model has some limitations our recommendations are the following:

* Incorporate more application questions related to default risk
* Add industry specific employment description
* Evaluate current economic conditions (such as recession) to applicant’s employment industry

## Introduction

Bad loans are one of the biggest hurdles to the development of the banking sector and have been increasingly studied and analyzed by researchers and authorities in recent years. This study shows the application of Logistic Regression for predictions if the loan will be fully repaid or not, and how bank make decisions about the eligibility of the loans.

The data set includes 30 variables of the loan for 50,000 loans.The data set contains quantitative and categorical variables and some values are missing. The purpose of this project is to build a plausible logistic regression model based on statistical analysis to predict good and bad loans and find a reasonable classification threshold between loans and profit for the bank. Each observation (record, row) represents one borrower and his/her information.

To predict loan defaults, first We need to identify the associations among the variables of the data set. Regression analysis starts with the loan information such as the interest rate, amount and income (Quantitative variables) to identify customers who did and did not pay the loan by its mature date (fully paid and charged off). Then combine this insight with information that would help identify the common characteristics of those customers, such as demographics, purpose, property information and employment (Categorical variables), that can then be used in the predictive model.

## Preparing and Cleaning the data

### Selection of variables for the analysis

loan <- loanData[c("grade","term","amount","rate","payment","status","length","home", "income","verified","reason","state","debtIncRat","delinq2yr","avgBal", "inq6mth","pubRec","openAcc","totalRevBal","revolRatio","totalPaid")]

### Eliminate variables

* employment - high amount of unique values
* accOpen24 - as we are using openAcc, we can eliminate this variable for using the predictive model
* totalAcc - total Account of the borrower (including closed accounts), will not use to predictive model as we are using open credit lines of the borrower.
* totalBal, totalLim - eliminating these 2 variables since we are using avgBal (average balance per account) for the predictor.
* bcRatio - there are 405 “NaN” values (missing Values) and there are 12062 “-Inf” values. Since there are high amount of missing values , this variable is not used for prediction model.
* totalBcLim and totalIlLim - I do not believe these 2 variables are critical factors for the predictor as we am using other variables related to credits.
* totalPaid - not a predictor. But this will remain in the data set as it will be used for test the model.

### Missing values and Data cleaning for missing observations

Observed missing values using the summary of the data set. (revolRatio had 19 missing values and other 11 variables had 1 missing value based on skim summary statistics.). Checked the missing values using is.na() function. Below are the solutions for the missing values of each variables.

**revolRatio** - assigned median value(0.55) for the missing values.

# checking summary of revolRatio to assign median value to missing values.  
summary(loan$revolRatio)

## Min. 1st Qu. Median Mean 3rd Qu. Max. NA's   
## 0.0000 0.3700 0.5500 0.5464 0.7300 1.6700 19

loan$revolRatio[is.na(loan$revolRatio)] <- 0.55

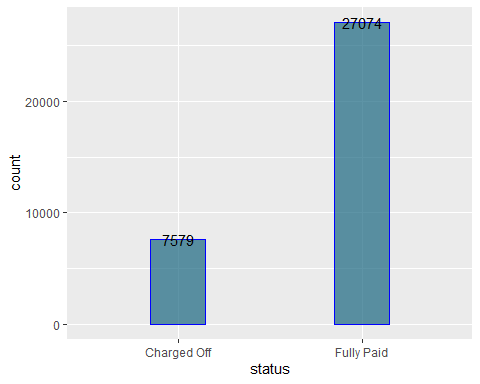
Other variables have only 1 missing value and these variables are important for the prediction model (eg: amount, payment, rate, income etc.), We can use complete.case function to return a logical vector indicating which cases are complete( have no missing values).

There are no missing values in the new loan vector.

### Variable - Status

Prepare the response variable based on the values of the status variable. Loans that are late, current (being paid), or in grace period have being removed from the data.

* Good Loans = fully paid, which means that the loan was paid
* Bad loans = loans with charged off and default status. Since there are only 2 loans in default status we will use only charged off status as bad loan, which means that there is no longer a reasonable expectation of further payments.

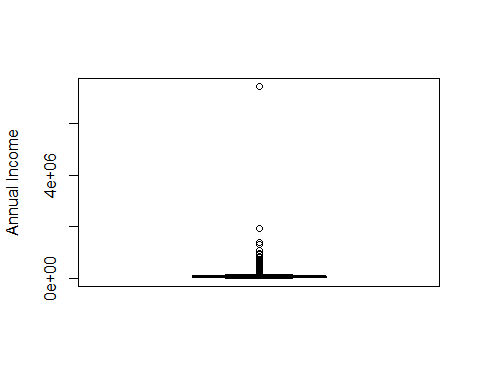


### Variable - Income

Creating scatter plot graph and summary of the income to identify outliers. There are potential outliers because 3rd quarter value is $90000 and the max(income) is $7446395.

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 4000 45500 65000 75220 90000 7446395

## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 1000 8000 12600 14621 20000 35000



In the box plot of annual income, we can see that there are some outliers. We would consider the average income, which lies between minimum income($4000) and $120000 (median US household income \* 2). It is highly unlikely that high income borrowers are included in the bad loan category and majority of population income falls within this range. The U.S. Census Bureau reported in September 2017 that real median household income was $59,039 in 2016(source-wikipedia). Based on loan amount summary, maximum loan amout is $35000, therefore it is reasonable to consider the borrowers with annual income less than or equal $120000.

### Variable - Length

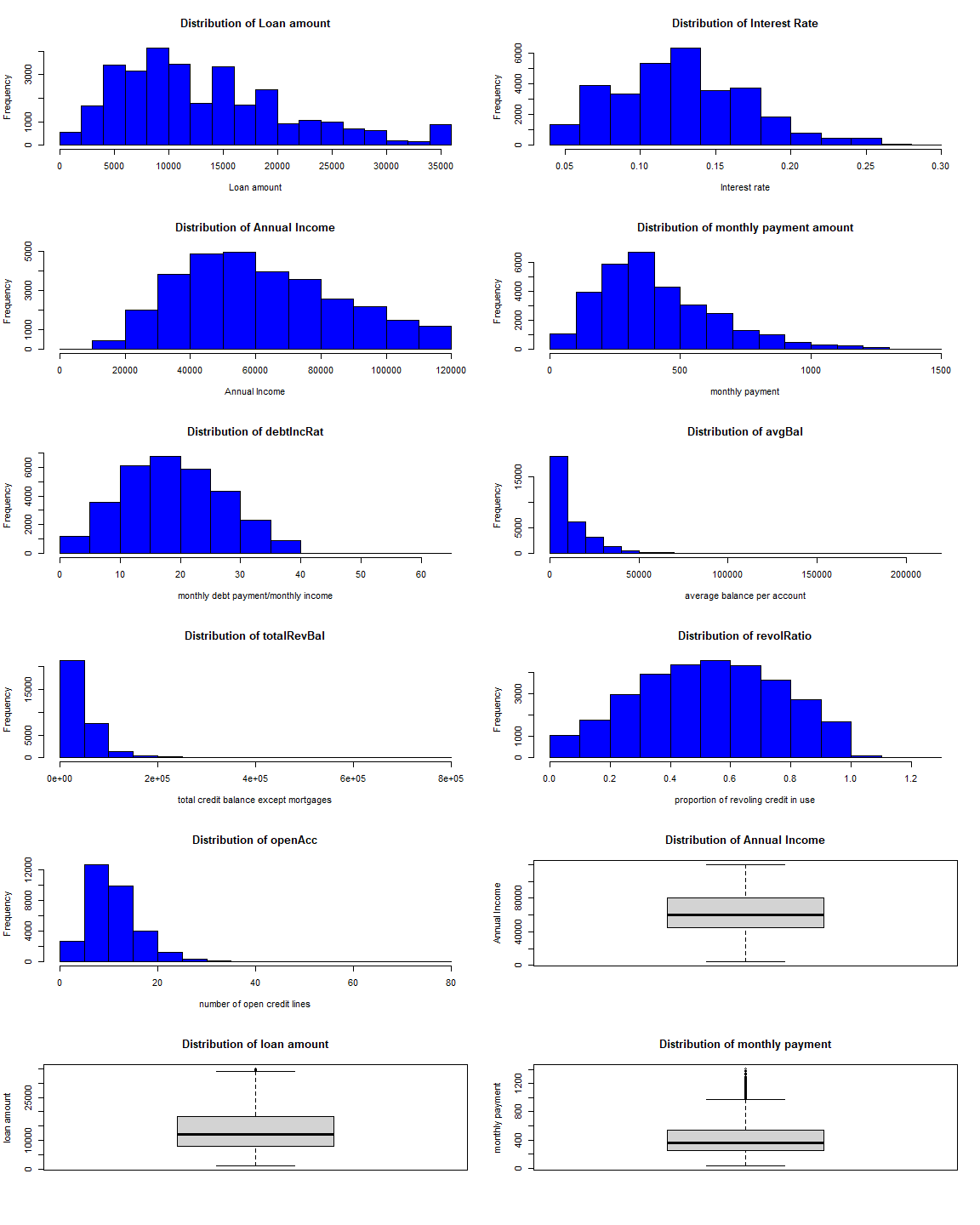
Observed values of employment length are “< 1 year”, “1 year”,“# years”, “10+ Years”, “n/a”. convert the “n/a” values to “1 year”. For analyzing purpose adding “< 1 year” to “1 year” and “10+ years” to “10 years”

### Other variables

Checking other variables for missing values and unique values using anyNA() and unique() functions. There are no missing values in the other variables.

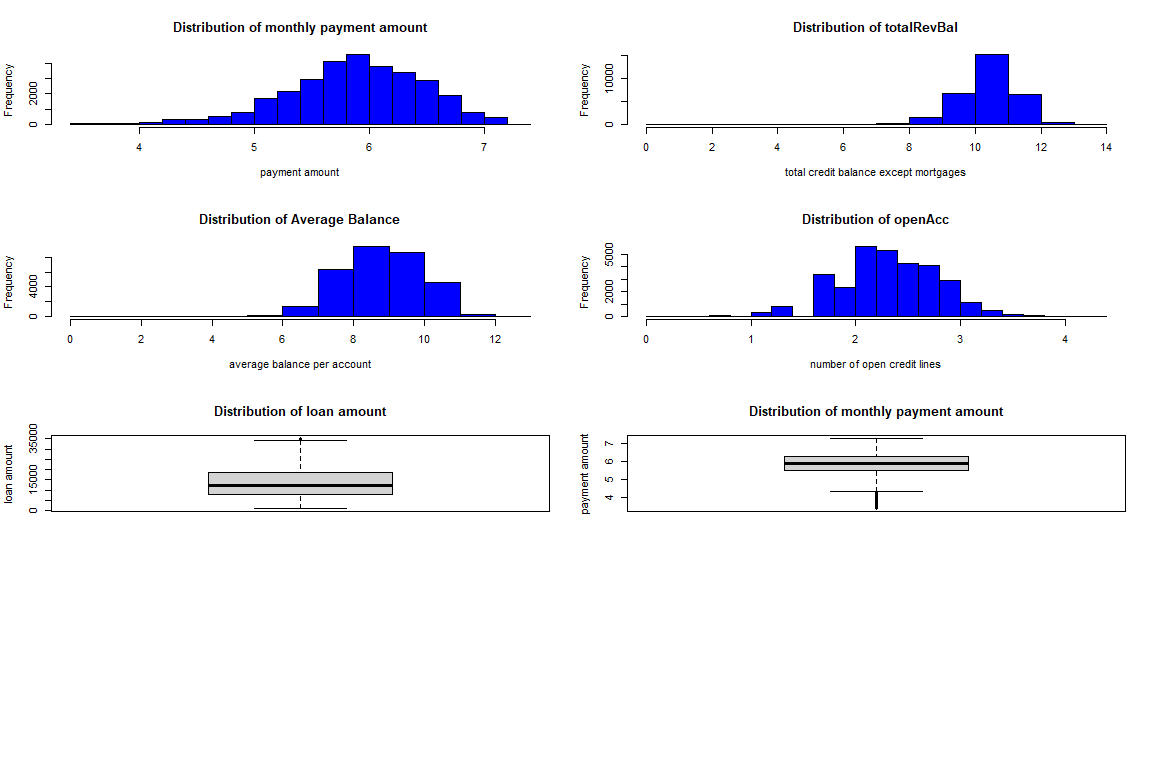
## Exploring and Transforming the data

### Distribution of the variables



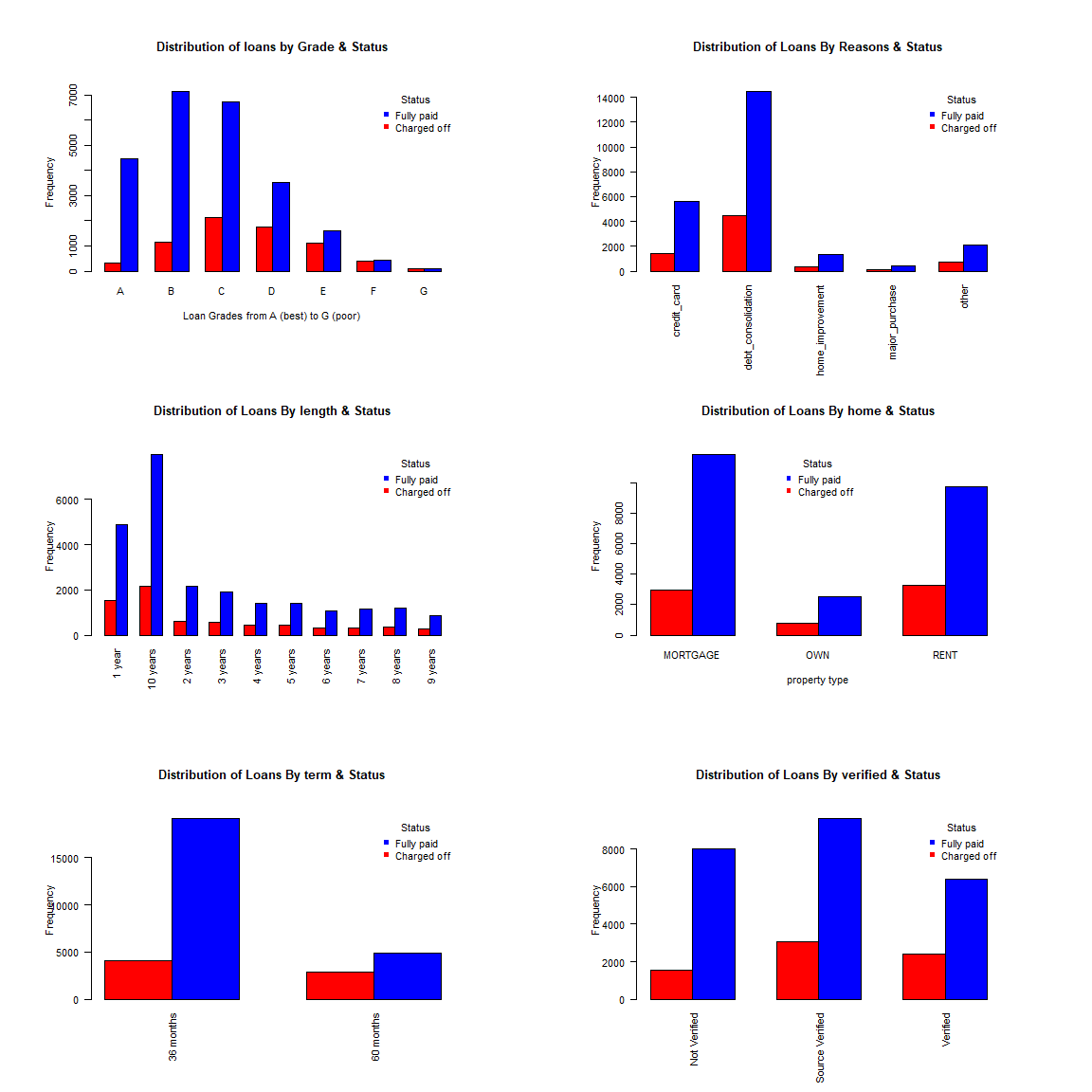
* Distribution of the interest rate, Annual income (box plot for the annual income does not show any outliers after transformation in the above section), debtIncRatio and revolRatio are approximately normal in shape.
* Distributions of the avgBal , and totalRevBal are highly skewed to right and needs to be transformed in order to use for predictive model.
* Distribution of openAcc does not look like normal in shape therefore will be transformed using logarithms transformation.
* There are few outliers in the box plot of loan amount and histogram shows the mildly skewed to the right. This variable will not be transformed as we will use it for profit calculation.
* There are many outliers in distribution of monthly payment amount and will be used log transformation to transform data.

### Data Transformation using logarithms transformation



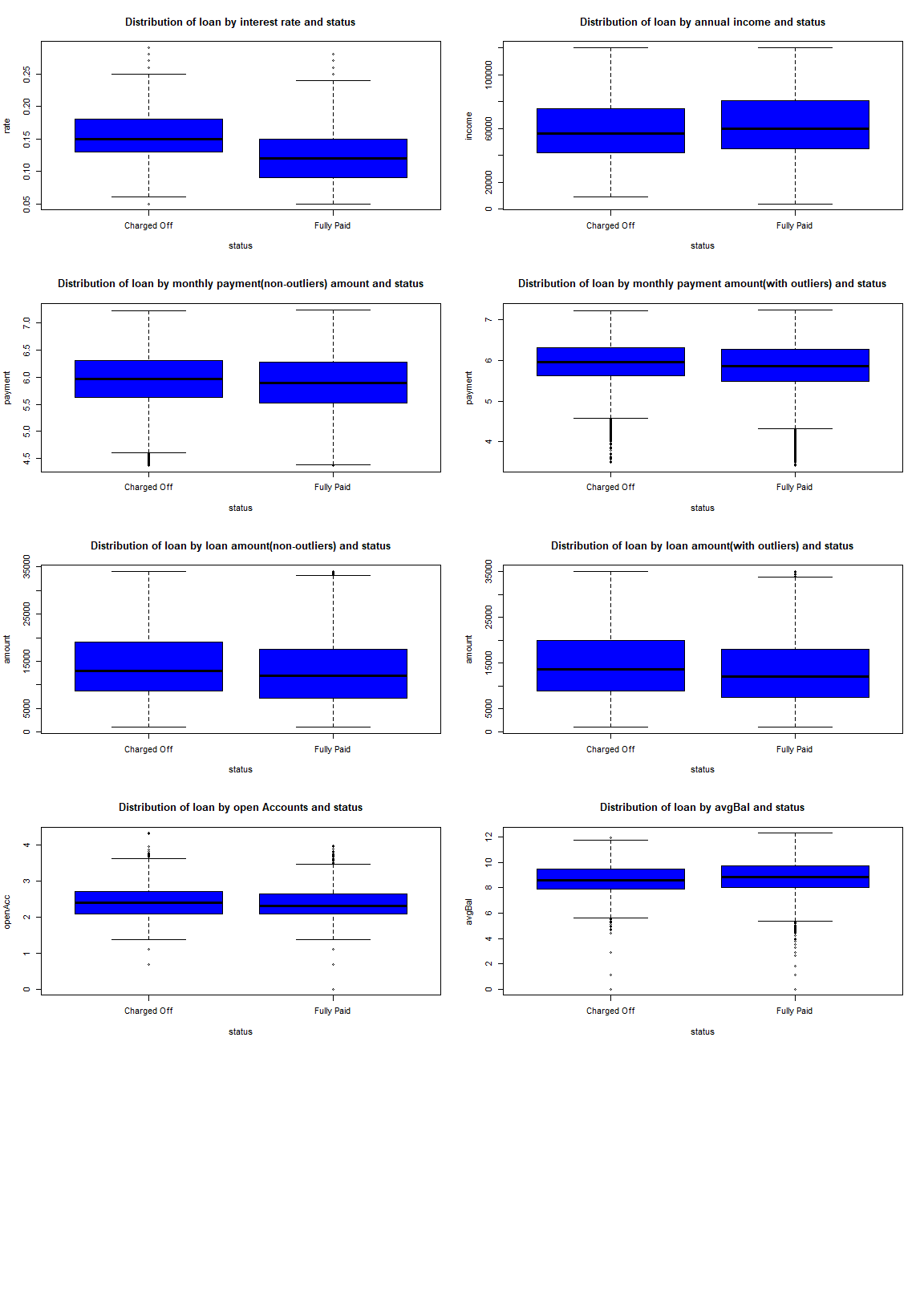
The logarithms transformation is a relatively strong transformation. It is a successful transformation for these 4 variables. While the transformed data here does not follow a normal distribution very well, it is probably about as close as we can get with these particular data.

## Distribution of Loans by categorical variables and loan status



There are 12 different categories for variable “reason”. Graph shows 4 most common reasons of taking loans by borrowers and combined all of the small categories into one larger “other” category. In the distribution of loan by grade score and status, we can see the percentage of unpaid loans increase as grade score quality decreases. Grade F has approximately similar number of charged-off borrowers and fully-paid borrowers and grade G has higher number of charged off borrowers. Therefore, we can conclude that loans grading has a direct relationship with the status. And also we can see home type “Rent” has higher percentage of charged off borrowers than “Mortgage” and “Own” property types. “60 months” term had higher percentage of charged off loans than “36 months”. Distribution of loan by verified and status behave quite differently for good and bad loans. surprisingly verified loans has higher percentage of charged off loans than not-verified loans.

## Distribution of Loans by quantitative variables and loan status



There are 12 different variable in this category. The graphs show the 6 most critical variables that can be used for predictive model. According to box plots, interest rate, monthly payment and loan amount are greater for the charged-off loans than fully-paid loan. The overall range of the data set is reasonably similar for the income, open accounts(charged-off borrowers median value is higher than paid-off borrowers) and average balance per account for charged-off and paid-off borrowers.

## The Logistic Model

### Creating two random samples for training and testing

## set the seed to make the partition reproducible  
set.seed(123)  
training\_sample <- sample(seq\_len(nrow(loan)), size = sample\_size)  
  
training <- loan[training\_sample, ]  
#drop totalPaid variable from training dataset  
train.data <- dplyr::select(training, -totalPaid, -outlier\_amount, -outlier\_payment)  
test.data <- loan[-training\_sample, ]

Before creating two random samples, converted loan status to binary variable, charged off loans as 0 and fully paid loans as 1. The next step is to create two sets of data out of loan data set; training data frame, containing 80% of loan sample and testing data frame with 20% of data for further modeling and analysis using sample() function. The “totalPaid” variable column is removed from the training data set and keep the “totalPaid” column in the test data to calculate the profit to find the optimize model for the profit maximization.

### Create a logistic regression model

Build the full logistic regression model using all of the predictors on training data to predict the loan status. The Full regression model is,

**status ~ grade + term + amount + rate + payment + length + home + income + verified + reason + state + debtIncRat + delinq2yr + avgBal + inq6mth + pubRec + openAcc + totalRevBal + revolRatio**

Then perform an ANOVA Chi-square test to check the overall effect of variables on the dependent variable (status) of the current model and see if there are any predictors with coefficients that do not significant at the 5% significance level. If there is no statistically significant relationship, which means no association between the change in the independent variable and the shifts in the dependent variable (Status). On the other hand these variables can make a negative effect for the accuracy of the model. Therefore based on the results we can see that the following variable should be dropped. payment, reason, pubRec, openAcc, totalRevBal and amount. However, we will consider variable “amount” as significant because it is a strong predictor in our model.

After dropping these variables, checking collinearity of the model by Calculating the variation inflation factors (VIF) of all predictors in regression model. The variables with high vif vales (>10) are “state” and “rate” and these 2 variables are dropped from the new model.

The new logistic regression model is,

**status ~ grade + term + amount + length + home + income + verified + debtIncRat + delinq2yr + avgBal + inq6mth + revolRatio**

### Determine the overall accuracy of the model

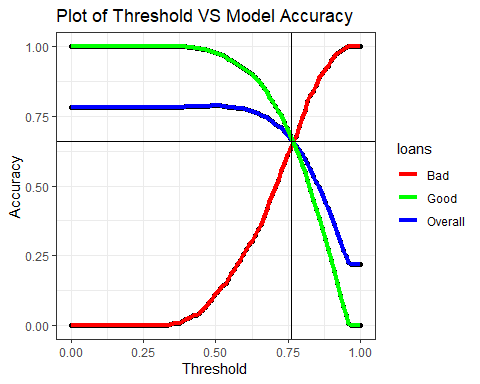
#Predicting outcomes on test data  
set.seed(123)  
  
#Cut-off value = 0.5  
predprob <- predict(training.logr2, newdata = test.data, type = "response")  
threshhold <- 0.5  
prediction <- cut(predprob, breaks=c(-Inf, threshhold, Inf),  
 labels=c("Bad", "Good"), header = TRUE)

In this section, predict the loan status for loans in the test data set using predict() function and determine the overall accuracy of the model at 0.5 threshold.

Based on the results produced, we can see that the model produced correctly predicted results of 78.72% and the percent of bad loans were correctly predicted as being bad is 10.81%. As for good loans we have had 98.11% predicted as good. We can note that with current model the correct prediction of the bad loan is very low and needs improvement.

### Optimizing the Threshold for Accuracy

After investigating optimum threshold to correctly predict both good and bad loans, threshold 0.8 can be used to predict the accuracy of the model. Based on the results produced we can note that the overall accuracy of the model have decreased from 78.72% to 60.07%, the percent of bad loans correctly predicted as bad have increased from 10.81% to 74.75% , and percent of good loans predicted as good have decreased from 98.11% to 55.88%. Even though our overall model accuracy slightly decreased but our prediction of bad and good loans have shown more balanced results for accuracy of the model; the results reflect that this model have become more effective of predicting both bad and good loans. In the next section, create graph showing the accuracy versus threshold to find the optimum threshold for the loan prediction and effect on overall accuracy.



We can consider optimum thershold for accuracy as the intercept of all three curves because it gives the more balanced accuracy levels for overall , good and bad loan predictions. The black vertical line shows the best threshold for accuracy in the regression model, which occurs at 0.76. The accuracy percentage of the model at best threshold for accuracy levels are,

* Overall accuracy of correctly predicted loans = 66.66 %"
* Accuracy of correctly predicted bad loans = 63.87 %"
* Accuracy of correctly predicted good loans = 67.44 %"

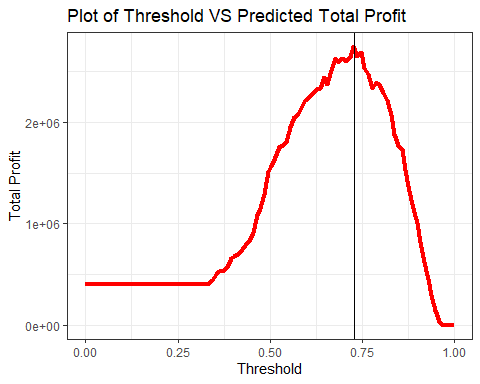
So far we were focused on how to improve our model by increasing the accuracy of the model prediction. From the bank’s perspective the most important feature of the model is to maximize overall profit. In the next section let’s examine the optimum threshold to maximize the profit.

### Accuracy of the model with optimum Threshold

Let’s look at the predicted profit result we can achieve by experimenting with the classification threshold; we will add “profit” column into our testing data frame by subtracting total amount repaid to the bank from the original loan amount column. After doing the prediction for the loan status using the model we will append the results to the testing data frame as “prediction”, and then calculate total profit based on predicted loan status.

## Optimizing the Threshold for Profit

If we use the same threshold found out in the above section as the best threshold to predict loan accuracy(0.76), the predicted profit would be a profit of $2,529,077 to the bank and the overall model accuracy is 66.66%, the model have correctly predicted 67.44% of good loans status, and bad loans predicted as 63.87%. We can see that this threshold is positively affects good loans prediction resulting in a significantly high profit for the bank. However, we believe the threshold can be improved a little more resulting in a more profit for the bank.



As a comparison for different levels of threshold affect on the total predicted profit we have created a scatter plot for visual comparison. We can see the threshold value of 0.73 is the most profitable threshold to the bank with overall model accuracy of 70.8%, the correctly predicted good loans of 74.76% versus incorrectly predicted of 25.24%. We are predicting the profit of $2,775,815 for the bank versus current calculated profit of $405,661 which results in a increase of profit by $2,370,154 (~2.37M).The percentage increase in profit that can be expected by deploying the model is,

print(paste('The percentage increase in total profit :',round(((sumprofitnew - sumprofit) / sumprofit) \* 100, digits = 2),'%'))

## [1] "The percentage increase in total profit : 584.27 %"

As we can see, the effect of threshold on the profit is very significant, however even setting the most profitable level for the bank, the trade off of the profit comes with the price of denying some of the loans that actually would have been good(incorrectly predicted good loans).

**Increase in profit compare to the increase in profit from a perfect model**

## [1] "The total profit from the perfect model: $ 10359695"

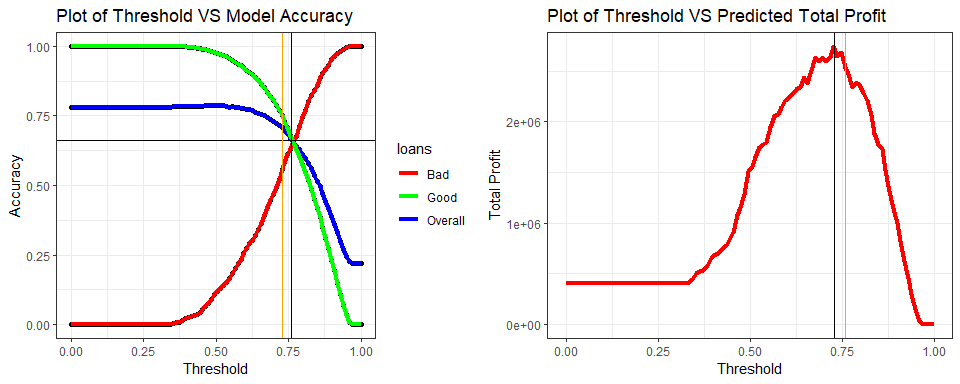
## [1] "The Percentage increase in the total profit when using Perfect Model VS our model : 8.98 %"

The perfect model would grant all loans that have a final status of “good” and deny all loans that have a final status of “bad”. Profit for the perfect model can be computed by filtering to include loans with actual final status “good” and then summing (totalPaid - loan amount). The total profit, the perfect model is generating approximately $10.4M and it is 8.98% increase of total profit than using our model.

**Overall accuracy and percentages of correctly predicted good and bad loans using best profit threshold (0.73)**

* Overall Accuracy of correctly predicted Loans = 70.8 %
* Accuracy of correctly predicted Bad Loans = 56.66 %
* Accuracy of correctly predicted Good Loans = 74.76 %

**Compare maximum profit threshold with the maximum accuracy threshold**



By using best profit threshold (orange line), we can see model accuracy of predicting overall and good loan has increased and the model accuracy of predicting bad loan has decreased. Accuracy difference when using best profit threshold, 0.73

* Overall : +4.14 %
* Good : +7.32 %
* Bad : -7.21%

By using optimum threshold for accuracy (orange line - 0.76), we can see there is a decrease in total profit, $246,738, which is equivalent to 8.9% decrease to maximum profit that can make with best profit threshold. Even though there is slightly decrease in profit, I do believe we should use optimum threshold for accuracy for predicting bad loans because the accuracy will be more important to build a good relationship with customers and ultimately increase the profit by increasing valuable customer base.

## Results Summary

This study provides building, statistical analysis, and evaluation of logistic regression model, including reasonable classification threshold in order to predict the loan status based on the loan application as well as predicted profit for the bank based on the suggested model. When building the model we have analyzed it in terms of correct prediction percent of fully paid and charged off loan’s status.

The determined model to maximize profit, provides an overall accuracy of 71% where correctly predicted fully paid loans are at 75% and charged off loans are at 57% and with a proposed threshold to maximize profit of 0.73 and the estimated predicted profit of $2.8M. However, the trade off of the profit decreased the correctly predicting charged off loans by about 7%, which can ultimately lose money to the bank.

We will suggest the model with optimum accuracy threshold (0.76), which has a capability of predicting the loans based on the information collected from applicants with overall accuracy of 67% where correctly predicted fully paid loans are at 67% and charged off loans are at 64%. With the proposed threshold of 0.76, the estimated predicted profit is $2.5M.By putting this model into action the bank can reduce the default loans and significantly increase the total profit.

However we believe the model can be improved even more if there will be created separate models for good and bad loan status prediction, possibly looking for more predictor variables that bank could collect, and create specific categories for the “employment” and “reason” variables. Even though we did drop these two variables (employment - high amount of unique values and reason - not significant) when doing the preparation, we do believe these variables would be helpful when running prediction.