Amity Institute of Integrative Sciences & Health (Manesar)

PROJECT REPORT

Predicting Loan Using Logistic Regression

Postgraduate Diploma in Data Science

Year 2020/2021

WALID FARAH DJAMA

Enrollment No : A52513450006

Summary

Executive Summary

Project objective

The main objective of the project, is to set out a good predictive Model, from the collected existing information on quality of the loan and improving the profit level with status of the loans, to set out the various variables of these standards and benchmarks to qualify for a good loans, and to extract from it a lists of minimum variables to predict a quality standards loans consideration for approval. The result is intended to improve a better revenue from the loans and be able to make profits while reducing the default loans in the process. The various elements of the project are summarised as follows:

- **Preparing and Cleaning the Data**
- **Exploring and Transforming the Data**
- **The Logistic Model**
- **Optimizing the Model for Accuracy**
- **Optimizing the Threshold for Profit**

Introduction

This project is about predicting loan and reducing defaults while improving profits using statistical analysis, and the evaluation of probable logistic regression model. I have dataset 50000 loans which is composed of 32 variables. And I have used R programming Languages for the analysis.

In my analysis have used several comprehensive aspect of generalized linear model in order to optimize and predict the loans status as well as predict profit that based on my model. I have analysed it in terms of correct prediction percent of fully paid and default loan's status and from there. I was able to predict a higher percentage of the profit and reduced the defaulted loans.

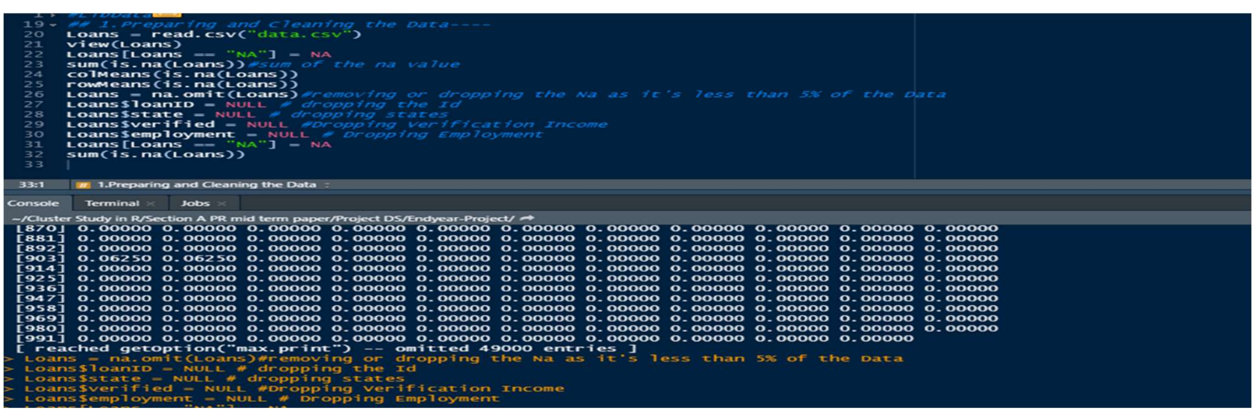
Preparing and Cleaning the Data

I have loaded all the library required for my analysis at first the loaded the dataset into the system.

Right after that I have started preparing/cleaning the data. I replaced original “n/a”, missing data, with the default value of NA for the purposes of my analysis. I usually impute missing data if its more than 5% but, in this case, the missing data is less than 5% so I had to drop some unnecessary variables from dataset as it will not have of any effect on my analysis. Our final dataset contains 28 variables instead of 32 variables.

Data Descriptions:

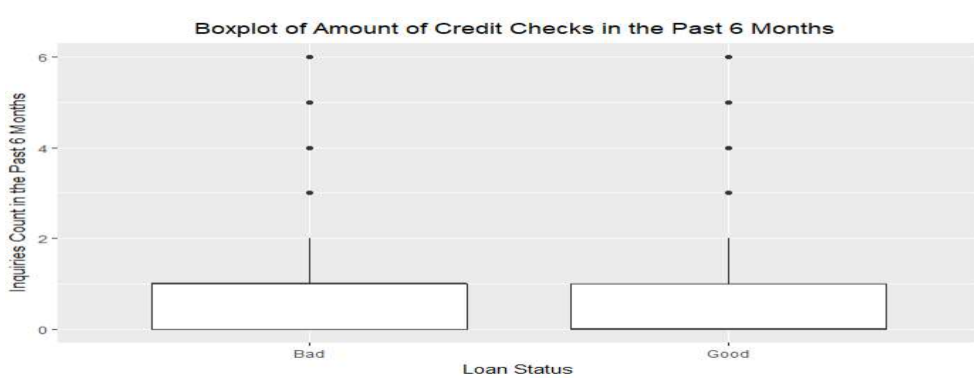
https://datascienceuwl.github.io/Project2018/TheData.html



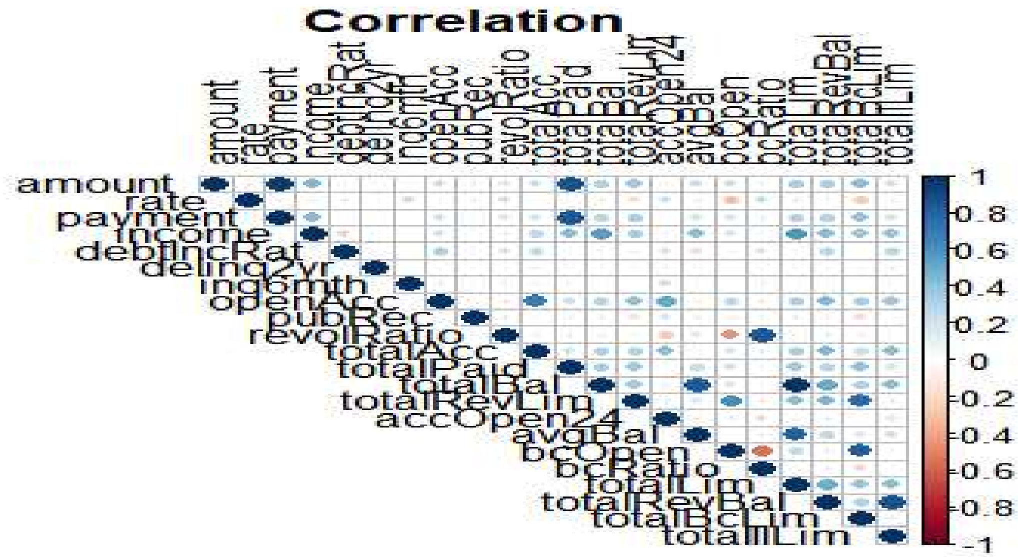
Exploring and Transforming the Data

I have started exploring and transforming the data as factor and characters and integer properly. The loan status contains multiple categories, for the purpose of my analysis I will retain only two categories of loan status: 'Fully Paid' a good loan, 'Charged Off' and 'Default' as bad loan status. We will also replace original "n/a", missing data, with the default value of NA for the purposes of my analysis.

After data preparation, we can now explore our dataset; first let's look at how many credit inquires have been done in the past 6 months for both good and bad loan status. As we can note on our boxplot, the median for Good loans is laying on 0 value meaning median inquiry for the past 6 months is no inquiries have been done, however for Bad loans status we see one inquiry; both boxplots are showing similar shape.



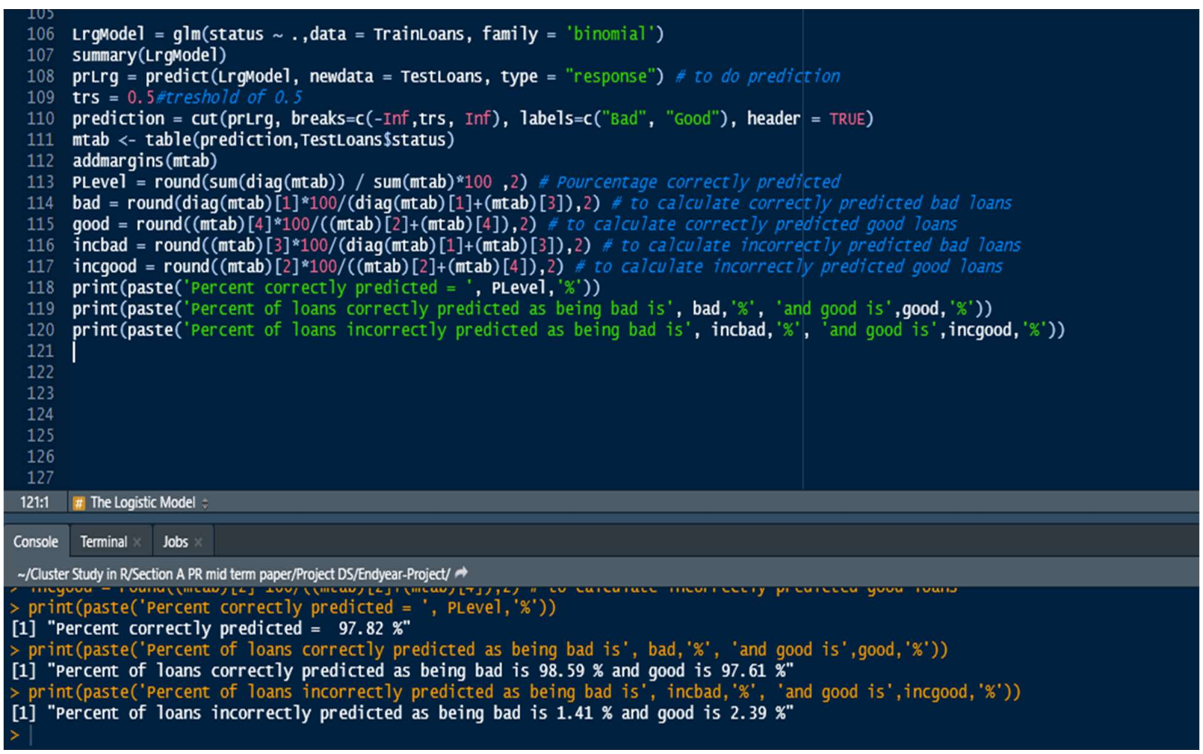
And then I checked the correlation between the variables and change the Loans status into categorical factor of Good and Bad Loans. And check the correlation using the Pearson method to see how they are correlated.




```
# 2.Exploring and Transforming the Data----
summary(Loans)
Profit = sum(Loans$amount-Loans$totalPaid)
Profit
Loans$status = as.factor(Loans$status)
levels(Loans$status) = list("Good"=c("Fully Paid"),"Bad"=c('Charged Off',"In Grace Period","Late (16-30 days)","Late (31-120 days)"))
table(Loans$status)
Loans$status = as.character(Loans$status)
Loans$length = as.character(Loans$length)
Loans$reason = as.character(Loans$reason)
Loans$amount = as.numeric(LoansInt$amount)
Loans$rate =  as.numeric(LoansInt$rate)
Loans$term = as.character.Date(Loans$term)
Loans$grade = as.character(Loans$grade)
Loans$length = as.character.Date(Loans$length)
2.Exploring and Transforming the Data                                          R Script
```

The Logistic Model

The logistic model that I used is generalized Linear model of a binomial family. The generalized linear model (GLM) is a flexible generalization of ordinary linear regression that allows for response variables that have error distribution models other than a normal distribution. So I used a glm algorithm where I used the factor of the as status against all other variable and the result was not satisfactory it was unbalanced as they were more good than bad in a greater value.

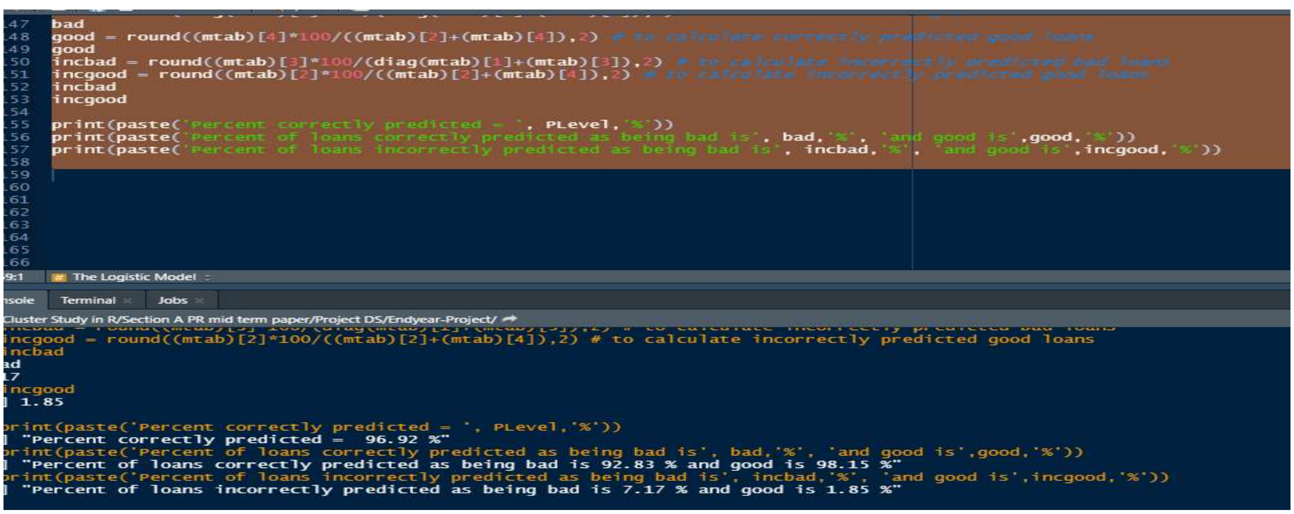


```r
LrgModel = glm(status ~ .,data = TrainLoans, family = 'binomial')
summary(LrgModel)
prLrg = predict(LrgModel, newdata = TestLoans, type = "response") # to do prediction
trs = 0.5#treshold of 0.5
prediction = cut(prLrg, breaks=c(-Inf,trs, Inf), labels=c("Bad", "Good"), header = TRUE)
mtab <- table(prediction,TestLoans$status)
addmargins(mtab)
PLevel = round(sum(diag(mtab)) / sum(mtab)*100 ,2) # Pourcentage correctly predicted
bad = round(diag(mtab)[1]*100/(diag(mtab)[1]+(mtab)[3]),2) # to calculate correctly predicted bad loans
good = round((mtab)[4]*100/((mtab)[2]+(mtab)[4]),2) # to calculate correctly predicted good loans
incbad = round((mtab)[3]*100/(diag(mtab)[1]+(mtab)[3]),2) # to calculate incorrectly predicted bad loans
incgood = round((mtab)[2]*100/((mtab)[2]+(mtab)[4]),2) # to calculate incorrectly predicted good loans
print(paste('Percent correctly predicted = ', PLevel,'%'))
print(paste('Percent of loans correctly predicted as being bad is', bad,'%', 'and good is',good,'%'))
print(paste('Percent of loans incorrectly predicted as being bad is', incbad,'%', 'and good is',incgood,'%'))
```

```
> print(paste('Percent correctly predicted = ', PLevel,'%'))
[1] "Percent correctly predicted =  97.82 %"
> print(paste('Percent of loans correctly predicted as being bad is', bad,'%', 'and good is',good,'%'))
[1] "Percent of loans correctly predicted as being bad is 98.59 % and good is 97.61 %"
> print(paste('Percent of loans incorrectly predicted as being bad is', incbad,'%', 'and good is',incgood,'%'))
[1] "Percent of loans incorrectly predicted as being bad is 1.41 % and good is 2.39 %"
>
```

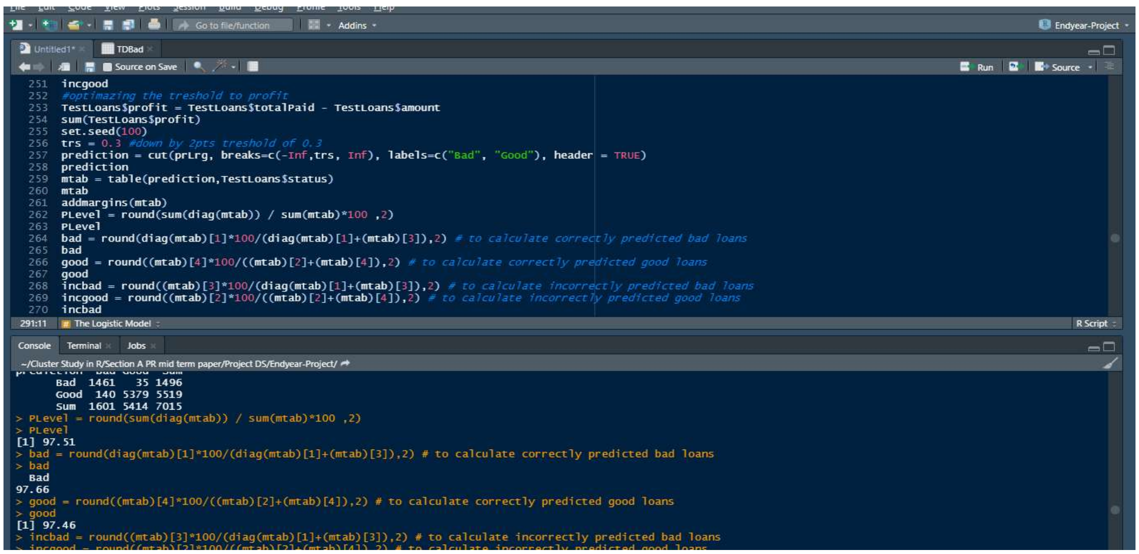
Optimizing the Threshold for Accuracy
(Balancing the data)

Based on the results produced, we can see that the full model produced correctly predicted results of 97.82% and the percent of bad loans were correctly predicted as being bad is 98.59% were correctly predicted). As for good loans we have had 97.61 %"predicted as good of correctly. I Can note that with current model the correct prediction of the bad loan is very higher than the good loan and needs improvement. So I used a new balanced and data the prediction as fallen a bit of Percent correctly predicted to 96.92 % but Percent of loans correctly predicted as being bad is 92.83 % and good is 98.15 %“ as good was good.



Optimizing the Threshold for Profit

I have optimized the threshold three model with 0.8 threshold the accuracy has dropped meanwhile at 0.3 it was far with 97.51 with considerable profit .



Results Summary

The determined model provides an overall accuracy where predicted correctly fully paid loans are at 93.26% with a proposed threshold of 0.3 with our estimated predicted profit of $ 12761340 However, the trade off of the profit comes with the price of denying some of the loans that actually would have been fully paid, the loan's status that was incorrectly predicted is 1.17%. The variation of the threshold can increase and decrease the percent of correctly predicted status, if we look at the graph of threshold level comparison to predicted profit on the right we will note that the profit is increasing up until threshold of about 0.3 and then decreasing as it goes up. Therefore, by implementing the proposed model the bank can increase their potential profit by $ 10855971.

Conclusion

So After performing data set cleaning and preparing as well as oversampling our training data set, then applying statistical model analysis we were able to build a plausible logistic regression model to predict good and bad loans and find a reasonable classification threshold in order to achieve the most profit for the bank. The model have shown the ability to predict the loans that will be fully paid off (Good) and charged off or default (Bad)

Therefore, we suggest for the bank prediction for the loan status is the following: status ~ totalPaid + amount + rate + payment + grade + bcRatio + debtIncRat + totalBal + length + term +totalAcc + openAcc + delinq2yr + totalBcLim + totalIlLim +totalRevBal + totalLim + totalRevLim + revolRatio + accOpen24 + bcOpen with 0.3 threshold applied.

I.e. : Though I dropped the employment variable when doing the preparation because of inconstancy , I do believe that a well classified employment variable would be helpful when running prediction.