Jordan Nushart

# Executive Summary

To improve loan selection processes, I created a logistic regression model to help us more accurately predict if a loan will be good or bad ahead of time. First, I started with a dataset of 50,000 records and 32 variables from prior loans and had to remove records and variables that were not useful. Once the data set was in its final form, I partitioned the data set into a training set that contained 80% of the remaining rows and a test set that contained the other 20% of the rows of data.

The training data set was the model that was used to make sure that we can build and train an accurate logistic regression model. This logistic regression model takes into consideration with strong effect the amount, payment, grade, length, whether they rent a home, the number of accounts opened last 24 months, debt-income ratio, delinquencies last 24 months, revolving ratio, and the total number of accounts. When I used the model from the training data on the test data, I was able to see the benefits of this model in the terms of a higher percentage of good loans given out and most importantly, increased profitability.

When the logistic regression model is used on the test data, we receive a score from 0-1 and can choose what we consider to be a loan worth giving. By changing the threshold of what we consider to be a good loan, we can be stricter giving out loans and optimize profit. When we leave the threshold at 0.50, we get the highest accuracy with good loans predicted at 80% accuracy, bad loans predicted with 60.6% accuracy, and combined accuracy of 79.2%.

Despite having the highest accuracy, the 0.50 threshold was not the ideal to have the highest profit. Our test data (6855 loans) had a profit of $1,654,791 under the current methods that the bank is using. With the logistic regression model and a 0.50 threshold, the profit on the test data would be $2,896,559.97 or a $1,241,768.97 increase over the current methods. If we become stricter and change the threshold to 0.67, we lower the accuracy to good loans predicted at 83.6% accuracy, bad loans predicted at 42.5% accuracy, and a combined accuracy of 75.3%. Despite the overall accuracy of the model going down, it was able to increase the profit in the test data to $3,754,097.88 or a $2,099,306.88 increase compared to the current methods. Multiple different thresholds were tested and compared the profits as follow (threshold|profit):  
0.35|$1,822,053.60; 0.40|$2,054,187.26; 0.45|$2,368,643.38; 0.50|$2,896,559.97; 0.55|$3,183,216.91; 0.60|$3,521,683.89; 0.65|$3,710,099.85; 0.67|$3,754,097.88

With the 0.67 threshold, the company was able to more than double the profits due to preventing loans that would be likely default or be charged off. By creating a model that focuses on the most important variables, we can more accurately determine what is a good loan to maximize profit. By using the logistic regression model with a 0.67 threshold, the bank can more than double its profits in the test data set and continue it with future loans and increase profitability going forward.

# Section 2 - Introduction

In this project, our goal is to be able to successfully predict if a loan will be good or bad given common information that would be available through a credit inquiry. The dataset that we are using has 32 variables, some of which we will remove. We will be using logistic regression and finding variables that have a difference between past good and bad loans to optimize profitability.

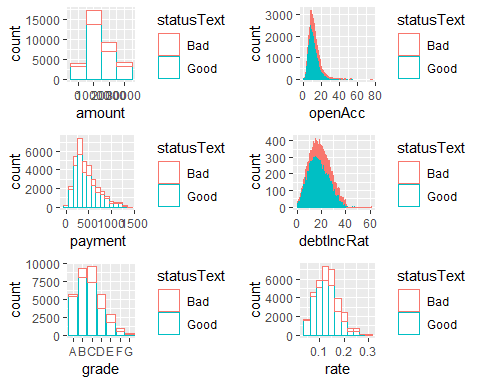
# Section 3 - Preparing and Cleaning the data

To prepare and clean the data, I started with removing a few columns that would be hard to group or use to predict if a loan would be good or not. Employment has too many unique entries to easily group. Meanwhile, Verified, and LoanID I do not believe would help predict if a loan would be successful. If entries had a length of employment that was ‘n/a’, I changed it to be ‘< 1 year’ since most entries did not have an entry for employment. I assume that most of these applicants were either unemployed or recently employed. I removed 521 rows that have n/a, the only variables left that have ‘n/a’ values are bcOpen and bcRatio and I did not believe that I could make a good assumption on what a ‘n/a’ could mean

I grouped states into 4 regions to make it easier to compare between regions. Next, I grouped the status variable into “Good” and “Bad” loans. Grouping into good and bad will allow for easy comparison when trying to find the most From there, I removed the other responses since the other loans are not completed so we cannot use them in factoring our model for good and bad loans. Filtering out the incomplete loans brings us down to 34,271 records remaining.

# Section 4 - Exploring and Transforming the data

Next, I looked at variables that I expected to be different between good and bad loans. In many of the graphs below, we can see a similar pattern where bad loans have more results at a higher amount for Open Acc, DebtIncRat, grade, rate, etc. My first instincts were that debt to income ratio, risk grade, and the interest rate would be the most important variables and from the histograms, they seem to be telling that the higher the applicant is, the more likely the loan is to default.



## $Bad  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 8796 43000 60000 68938 82000 1300000   
##   
## $Good  
## Min. 1st Qu. Median Mean 3rd Qu. Max.   
## 7067 47000 65000 76968 92700 7446395

# Section 5 - The Logistic Model

Here I took 80% of the rows that were left in the dataset and created a dataset to train a logistic regression model to use on the other 20% of the dataset. When looking at the summary of the logistic regression model, I was able to see which variables are most important to help us predict the correct outcome. When looking at the output below, I tested many models trying to find which one had the lowest AIC value which would mean that it is the most accurate. All variables that did not have a response with a pr value under 0.05 were removed to help make the model more accurate.

## training test   
## 0.7999767 0.2000233

## Analysis of Deviance Table  
##   
## Model: binomial, link: logit  
##   
## Response: status  
##   
## Terms added sequentially (first to last)  
##   
##   
## Df Deviance Resid. Df Resid. Dev Pr(>Chi)   
## NULL 27415 28850   
## amount 1 68.23 27414 28781 < 2.2e-16 \*\*\*  
## term 1 971.68 27413 27810 < 2.2e-16 \*\*\*  
## rate 1 1227.40 27412 26582 < 2.2e-16 \*\*\*  
## payment 1 10.97 27411 26571 0.0009246 \*\*\*  
## grade 6 168.59 27405 26403 < 2.2e-16 \*\*\*  
## length 11 70.38 27394 26332 1.035e-10 \*\*\*  
## home 2 113.45 27392 26219 < 2.2e-16 \*\*\*  
## income 1 27.08 27391 26192 1.949e-07 \*\*\*  
## reason 12 19.85 27379 26172 0.0699863 .   
## debtIncRat 1 110.75 27378 26061 < 2.2e-16 \*\*\*  
## delinq2yr 1 6.55 27377 26055 0.0104839 \*   
## inq6mth 1 17.15 27376 26038 3.455e-05 \*\*\*  
## openAcc 1 6.37 27375 26031 0.0116272 \*   
## pubRec 1 2.85 27374 26028 0.0910912 .   
## revolRatio 1 4.60 27373 26024 0.0318894 \*   
## totalAcc 1 26.34 27372 25997 2.869e-07 \*\*\*  
## totalBal 1 48.06 27371 25949 4.135e-12 \*\*\*  
## totalRevLim 1 8.71 27370 25941 0.0031665 \*\*   
## accOpen24 1 163.81 27369 25777 < 2.2e-16 \*\*\*  
## bcOpen 1 9.32 27368 25767 0.0022710 \*\*   
## bcRatio 1 4.92 27367 25763 0.0265912 \*   
## totalLim 1 2.98 27366 25760 0.0843677 .   
## totalBcLim 1 3.67 27365 25756 0.0555559 .   
## totalIlLim 1 5.02 27364 25751 0.0250952 \*   
## region 3 27.10 27361 25724 5.619e-06 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1

##   
## Call:  
## glm(formula = status ~ ., family = "binomial", data = training2)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9598 0.2946 0.5133 0.7299 1.8720   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.124e+00 2.517e-01 12.414 < 2e-16 \*\*\*  
## amount 3.723e-05 1.383e-05 2.692 0.007103 \*\*   
## term 60 months -9.206e-01 8.921e-02 -10.319 < 2e-16 \*\*\*  
## rate -7.230e-01 1.391e+00 -0.520 0.603329   
## payment -1.551e-03 4.336e-04 -3.577 0.000348 \*\*\*  
## gradeB -4.564e-01 8.394e-02 -5.437 5.43e-08 \*\*\*  
## gradeC -8.638e-01 1.100e-01 -7.855 4.01e-15 \*\*\*  
## gradeD -1.070e+00 1.503e-01 -7.119 1.09e-12 \*\*\*  
## gradeE -1.138e+00 1.853e-01 -6.140 8.25e-10 \*\*\*  
## gradeF -1.241e+00 2.456e-01 -5.053 4.35e-07 \*\*\*  
## gradeG -1.343e+00 3.095e-01 -4.338 1.44e-05 \*\*\*  
## length10 4.619e-01 9.487e-02 4.869 1.12e-06 \*\*\*  
## length11 4.206e-01 9.186e-02 4.579 4.68e-06 \*\*\*  
## length12 4.291e-01 1.025e-01 4.187 2.83e-05 \*\*\*  
## length2 4.395e-01 8.314e-02 5.286 1.25e-07 \*\*\*  
## length3 5.006e-01 8.731e-02 5.734 9.81e-09 \*\*\*  
## length4 5.289e-01 6.848e-02 7.723 1.14e-14 \*\*\*  
## length5 4.795e-01 8.188e-02 5.857 4.71e-09 \*\*\*  
## length6 4.549e-01 8.334e-02 5.459 4.80e-08 \*\*\*  
## length7 4.278e-01 8.893e-02 4.810 1.51e-06 \*\*\*  
## length8 4.417e-01 8.852e-02 4.989 6.06e-07 \*\*\*  
## length9 4.156e-01 9.529e-02 4.361 1.29e-05 \*\*\*  
## homeOWN -6.625e-02 5.595e-02 -1.184 0.236426   
## homeRENT -2.685e-01 4.193e-02 -6.404 1.51e-10 \*\*\*  
## income 1.147e-09 5.205e-07 0.002 0.998243   
## reasoncredit\_card -3.383e-01 2.007e-01 -1.686 0.091829 .   
## reasondebt\_consolidation -2.809e-01 1.982e-01 -1.417 0.156461   
## reasonhome\_improvement -4.195e-01 2.079e-01 -2.018 0.043600 \*   
## reasonhouse -3.316e-01 3.109e-01 -1.067 0.286158   
## reasonmajor\_purchase -3.961e-01 2.292e-01 -1.728 0.084014 .   
## reasonmedical -6.296e-01 2.401e-01 -2.623 0.008726 \*\*   
## reasonmoving -7.857e-01 2.671e-01 -2.942 0.003266 \*\*   
## reasonother -2.951e-01 2.094e-01 -1.409 0.158732   
## reasonrenewable\_energy -4.253e-01 5.253e-01 -0.810 0.418128   
## reasonsmall\_business -7.550e-01 2.420e-01 -3.120 0.001811 \*\*   
## reasonvacation -4.389e-01 2.752e-01 -1.595 0.110784   
## reasonwedding 7.983e+00 7.246e+01 0.110 0.912279   
## debtIncRat -2.257e-02 2.372e-03 -9.515 < 2e-16 \*\*\*  
## delinq2yr -8.651e-02 1.729e-02 -5.004 5.61e-07 \*\*\*  
## inq6mth -5.070e-02 1.614e-02 -3.141 0.001686 \*\*   
## openAcc -1.369e-02 4.407e-03 -3.107 0.001888 \*\*   
## pubRec -1.950e-02 2.473e-02 -0.789 0.430356   
## revolRatio -5.113e-01 1.242e-01 -4.116 3.86e-05 \*\*\*  
## totalAcc 1.249e-02 1.926e-03 6.484 8.93e-11 \*\*\*  
## totalBal -9.606e-07 9.305e-07 -1.032 0.301885   
## totalRevLim -9.759e-07 9.327e-07 -1.046 0.295380   
## accOpen24 -7.510e-02 6.132e-03 -12.247 < 2e-16 \*\*\*  
## bcOpen 1.215e-06 3.388e-06 0.359 0.719866   
## bcRatio 1.979e-03 1.079e-03 1.833 0.066740 .   
## totalLim 1.754e-06 8.826e-07 1.987 0.046904 \*   
## totalBcLim 4.868e-06 2.194e-06 2.219 0.026516 \*   
## totalIlLim 1.218e-06 5.337e-07 2.282 0.022474 \*   
## regionNortheast -8.722e-02 5.100e-02 -1.710 0.087230 .   
## regionSouth -2.346e-02 4.551e-02 -0.516 0.606200   
## regionWest 1.389e-01 4.925e-02 2.820 0.004807 \*\*   
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 28850 on 27415 degrees of freedom  
## Residual deviance: 25724 on 27361 degrees of freedom  
## AIC: 25834  
##   
## Number of Fisher Scoring iterations: 8

## predLoan  
## 0 1 Sum  
## 0 175 1314 1489  
## 1 114 5252 5366  
## Sum 289 6566 6855

## [1] "Accuracy for good loans is : 0.799878160219312"

## [1] "Accuracy for bad loans is : 0.605536332179931"

## [1] "Combined accuracy is : 0.791684901531729"

# Section 6 - Optimizing the Threshold for Accuracy

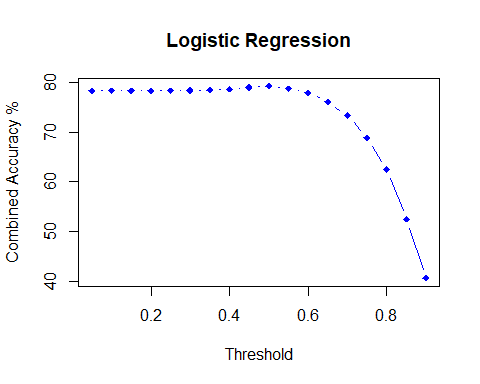
Next, we compared the predicted results of the model against the actual results of the loan to determine how accurate the model is. From the output below, we can see that the combined accuracy is around 79.2% when using a threshold value of 0.5 and higher for a ‘Good’ result from a loan. Good loans were predicted with 80% accuracy and bad loans were predicted with 60.6% accuracy but there are far fewer bad loans than good loans. I created a chart to show the accuracy percentage at every 0.05 increase in the threshold. 0.5 threshold was the most accurate in this test dataset.

## predLoan  
## 0 1 Sum  
## 0 591 898 1489  
## 1 798 4568 5366  
## Sum 1389 5466 6855

## [1] "Accuracy for good loans is : 0.835711672155141"

## [1] "Accuracy for bad loans is : 0.42548596112311"

## [1] "Combined accuracy is : 0.752589350838804"



# Section 7 - “Optimizing the Threshold for Profit

Then, I focused on the model output to predict how much profit the bank would see. I calculated the profit/loss for the current data set. I took the predicted value, compared that to the threshold, gave the predicted value a ‘Good’ or ‘Bad’ rating, and calculated profit/loss for all of the predicted Good loans. Here, I tested multiple thresholds to determine which threshold would give the bank the highest profits and percentage increase in profits compared to the test data set. As you can see below, I tested many thresholds from 0.5 to the optimal threshold of 0.67. With a threshold of 0.67, the bank would’ve had a profit of $3,754,097.88 which is around 127% higher than the profit the bank saw on their loans using the current method ($1,654,791).

## [1] "threshold = 0.35 profit = 1822053.6 0.1011"

## [1] "threshold = 0.4 profit = 2054187.26 0.2414"

## [1] "threshold = 0.45 profit = 2368643.38 0.4314"

## [1] "threshold = 0.5 profit = 2896559.97 0.7504"

## [1] "threshold = 0.55 profit = 3183216.91 0.9236"

## [1] "threshold = 0.6 profit = 3521683.89 1.1282"

## [1] "threshold = 0.65 profit = 3710099.85 1.242"

## [1] "threshold = 0.67 profit = 3754097.88 1.2686"

# Section 8 - Results Summary

The logistic regression model shows great potential to improve the bank’s loan processes. By removing variables that were not significant in determining whether a loan would be good or bad, we were able to create an effective logistic regression model that provided great results. This logistic regression model takes into consideration with strong effect the amount, payment, grade, length, whether they rent a home, the number of accounts open in the last 24 months, debt-income ratio, delinquencies in the last 24 months, revolving ratio, the total number of accounts, and accounts opened in the last 24 months. This modelThe following statistics are from results used only on the test dataset which is 20% of the total dataset given to us:  
-With a 0.50 threshold, the model will predict the correct outcome of a loan around 79.2% of the time.  
+With the 79.2% accuracy, the bank would earn a profit of $2,896,559.97, which would be an increase over the current approval methods by $1,241,768.97, or 75.04%.  
-My recommendation is to instead use a 0.67 threshold which would drop the accuracy down around 75.3% but would increase the profits to $3,754,097.88.  
+This would be a 127% or $2,099,306.88 increase in profits compared to the current methods.