DS 705 Project Part 3 Final Report

Korey Bernhardt

4/28/2020

## Executive Summary

This report will examine and recommend criteria to help determine the likelihood of loan applicants defaulting on a loan in an effort to improve accuracy and profitability of approving loans to applicants.

Twenty eight different factors that are available on loan applicants were examined, and a model was created based on their relationship to positive or negative outcomes of existing loans. The goal was to improve upon the model currently in use in both accuracy and profit, recognizing that there is a point at which improving the potential accuracy on a forecasting model may actually result in less profit. In other words, if you decline every loan applicant you will never approve a bad loan (100% accurate), but you also won’t realize any profits.

Using the model created we were able to achieve an accuracy level of between 79.05% and 74.6%, and increase in profits of between 116.74% and 253.52%. The factors that are used to create this model include:

|  |  |
| --- | --- |
| Factors |  |
| Interest rate | Credit limits for installment accounts |
| Loan term (term) | 30+ day late payments in past 2 years |
| Average balance per account | Prop of revolving credit in use |
| Debt to income ratio | Annual income verification status |
| Accounts opened in past 2 years | Credit checks in past 6 months |
| Total number of credit lines | Monthly payment amount |
| Home ownership | Credit limits of credit cards |
| Grade of loan |  |

I recommended beginning using this model instead of the current one given the potential for improvement of both accuracy and profit for the bank. This model and future results should also be reviewed on a regular basis to determine how well the model is performing, and if changes should be made.

## Introduction

What factors can help determine risk and maximize profits from loans made by banks? Using a dataset with variables such as loan amount, length of employment, home ownership, income, revolving credit in use and accounts opened in the past 2 years we’ll explore what variables are most correlated to a good or bad outcome for the bank from a loan, and develop a model to improve the accuracy and profitability of loan approvals.

## Preparing and Cleaning the data

The initial data set included 50,000 rows and 32 variables. In order to determine if a loan had a good or a bad outcome, a status field was used to interpret loans with a status of Fully Paid as good loans, and those that had a status of Charged Off as bad loans. Data that had any other status was removed as it didn’t indicate an outcome. 1 row was also removed as it had primarily invalid data (NA) for a majority of the variables. This left 34,655 rows of data. Four variables were also removed from the data set as they were not thought to be predictor variables. These were the loan ID, job title, state, and total amount paid.

In order to determine if the variables in the data had a correlation to status, several of the variables were adjusted to be numeric and in some cases grouped together to provide more meaningful references. Status, term, home ownership and whether income was verified were adjusted to be numeric. The following groupings were performed:

* *Grade*: A and B were grouped together, C and D were grouped together, and F was it’s own group. These represented a low, medium, and high risk categorization.
* *Length of continuous employment*: less than 1 through 3 years were grouped together, 4 through 9 years were grouped together, and 10+ was it’s own group. These groupings were primarily based on volume of the groupings, so that it wouldn’t be a larger volume of loans creating a stronger correlation.
* *Reason for the loan*: There were 2 options that were significantly larger than all others combined, credit card and debt consolidation. These 2 were left separate, and all other reasons were combined to form an “other” category.

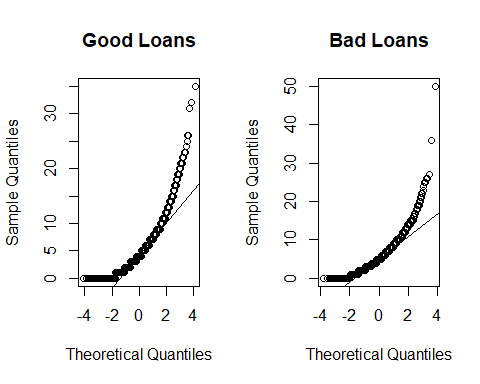
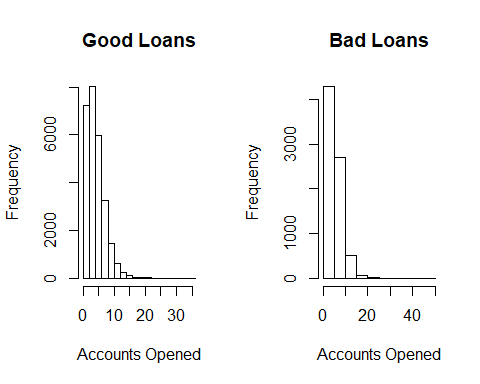
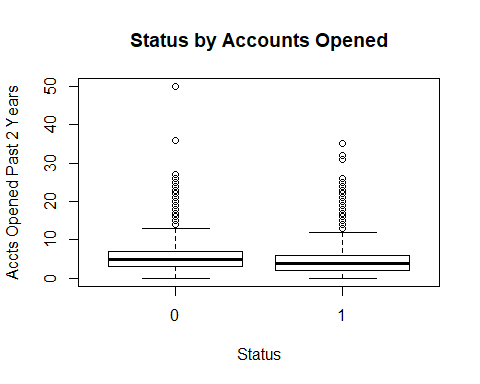
The remaining 28 variables were each compared to status in order to determine the coefficient to help determine which would be most likely to provide a correlation to the outcome of the loan. The top 5 correlations with status were term, rate, grade, debt to income ratio and accounts opened in the past 24 months. The results of the analysis for these 5 variables is shown below.

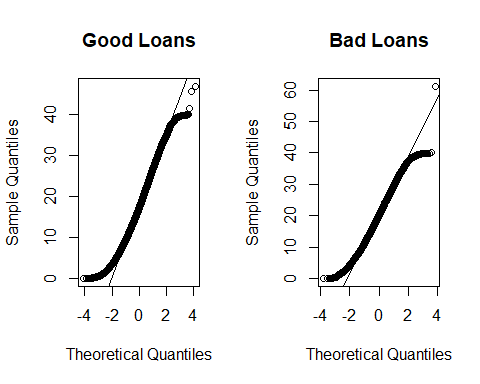
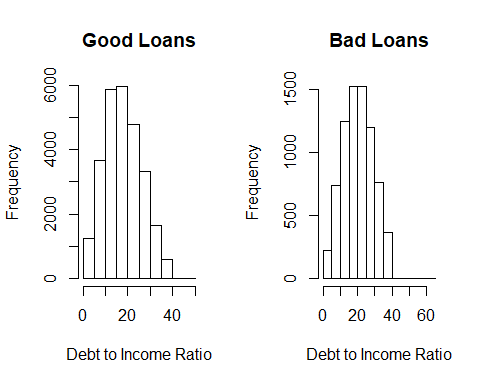
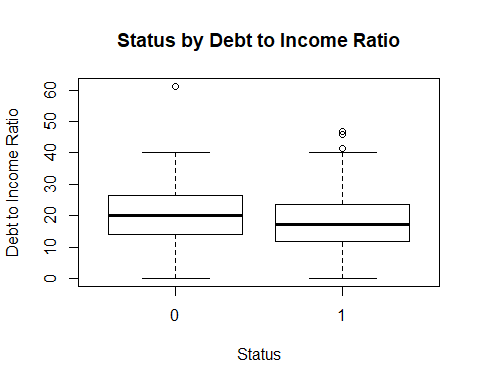
## Variable1 Variable2 Coeffiecient  
## 1 status status\_factor 1.0000000  
## 2 totalPaid status\_factor 0.3547449  
## 3 rate status\_factor -0.2740104  
## 4 grade status\_factor -0.2621182  
## 5 term status\_factor -0.1996039

## Exploring and Transforming the data

In further exploring the data, I chose to target in on the number of accounts opened in the past 2 years and Debt to Income Ratio. While there may be some connection between these, out of the top 5 variables the others are less independent than these.

Both of these variables, as predictors of loan status, appear to be normally distributed but have outliers and some skewing. In order to overcome this, I converted them each to log values and square roots to do further comparison. This didn’t improve the normality of either variable.





## The Logistic Model

The target response variable for the prediction model is status. In order to evaluate potential models, forward step, backward step and regsubsets approaches were investigated. Further model fitting included removing variables that had high collinearity and removing insignificant P-values. Based on AIC and adjusted R2 comparisons, a model resulting from forward step creation was selected. The variables used in this model are: rate, term , avgBal, debtIncRat, accOpen24, totalAcc, home, grade, totalLim, delinq2yr, revolRatio, verified, inq6mth, payment and totalBcLim.

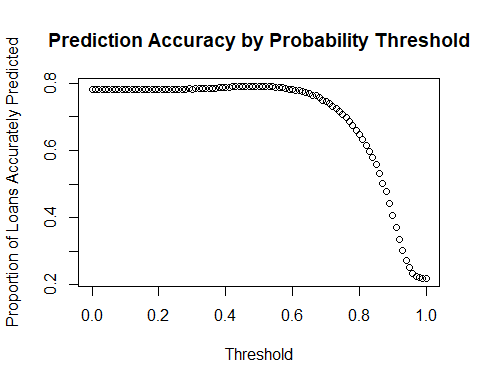
Variables that were removed due to collinearity were: amount and totalRevBal. Variables removed due to insignificant P-values were: bcOpen, totalIlLim and totalRevLim.

The final model is summarized below. 20% of the original data set containing completed loans (those that had been either fully paid or charged off) was reserved for testing the model, which resulted in 78.9% of the loan statuses being accurately predicted at a probability cutoff of 0.5.

##   
## Call:  
## glm(formula = status ~ rate + term + avgBal + debtIncRat + accOpen24 +   
## totalAcc + home + grade + totalLim + delinq2yr + revolRatio +   
## verified + inq6mth + payment + totalBcLim, family = "binomial",   
## data = training\_loans)  
##   
## Deviance Residuals:   
## Min 1Q Median 3Q Max   
## -2.9077 0.3392 0.5272 0.7138 1.8248   
##   
## Coefficients:  
## Estimate Std. Error z value Pr(>|z|)   
## (Intercept) 3.945e+00 1.037e-01 38.022 < 2e-16 \*\*\*  
## rate -5.783e+00 7.697e-01 -7.514 5.76e-14 \*\*\*  
## term -2.681e-02 1.582e-03 -16.940 < 2e-16 \*\*\*  
## avgBal 7.637e-06 2.494e-06 3.062 0.002195 \*\*   
## debtIncRat -2.232e-02 1.958e-03 -11.399 < 2e-16 \*\*\*  
## accOpen24 -7.447e-02 5.798e-03 -12.844 < 2e-16 \*\*\*  
## totalAcc 1.158e-02 1.649e-03 7.023 2.17e-12 \*\*\*  
## home 1.196e-01 2.549e-02 4.692 2.70e-06 \*\*\*  
## grade -2.165e-01 4.730e-02 -4.577 4.72e-06 \*\*\*  
## totalLim 6.353e-07 2.412e-07 2.634 0.008427 \*\*   
## delinq2yr -7.738e-02 1.753e-02 -4.415 1.01e-05 \*\*\*  
## revolRatio -3.264e-01 7.313e-02 -4.464 8.06e-06 \*\*\*  
## verified -6.081e-02 2.175e-02 -2.796 0.005175 \*\*   
## inq6mth -5.186e-02 1.594e-02 -3.253 0.001143 \*\*   
## payment -1.823e-04 7.387e-05 -2.468 0.013573 \*   
## totalBcLim 4.241e-06 1.096e-06 3.869 0.000109 \*\*\*  
## ---  
## Signif. codes: 0 '\*\*\*' 0.001 '\*\*' 0.01 '\*' 0.05 '.' 0.1 ' ' 1  
##   
## (Dispersion parameter for binomial family taken to be 1)  
##   
## Null deviance: 29124 on 27723 degrees of freedom  
## Residual deviance: 26288 on 27708 degrees of freedom  
## AIC: 26320  
##   
## Number of Fisher Scoring iterations: 4

## Optimizing the Threshold for Accuracy

In order to evaluate a probability threshold to use to improve the accuracy, the threshold was varied from 0 to 1 by increments of 0.01. The threshold that produced the highest accuracy was 0.45, resulting in 79.05% of the loans being accurately predicted in the test data. The accuracy is fairly stable up to about a 0.75 probability where it begins to fall more quickly.



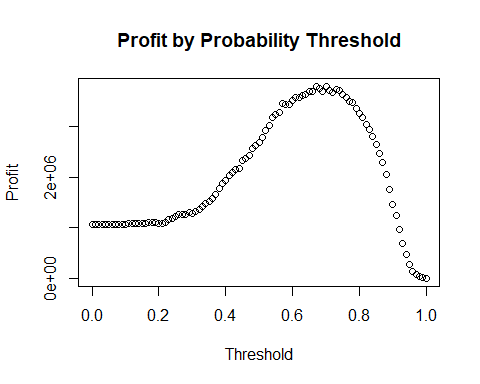
## Optimizing the Threshold for Profit

In order to evaluate a probability threshold to use to improve the profit level, the threshold was again varied from 0 to 1 by increments of 0.01. The threshold that produced the highest profit was 0.70, resulting in a profit of $3,802,714 from loans in the test data. This is quite a bit higher than the best accuracy threshold of 0.45, which would result in a profit of $2,331,409. This threshold is also close to the point at which the accuracy of the model begins to rapidly decline.

When comparing the model to the actual outcome of the loans, the model produced a profit that was 253.52% more than the actual results that would have been produced of a profit of $1,075,686 from the test data.

Alternatively, a perfect outcome of only approving loans that result in fully paid outcomes would have resulted in a profit of $12,447,264 which is 1057.15% more than the actual results based on the test data. That said, the model still produces a significantly higher profit potential than the actual approach used.

The accuracy at the 0.70 threshold used to obtain the highest profit is 74.6%, so there is some risk at that threshold that the model may not perform as well and ultimately result in less profits than we’re seeing with the test data.



## Results Summary

After selecting a model and evaluating the results, this model does provide a significant improvement over the current approach to approving loans. Applying a probability threshold of between 0.45 and .70 would result in an accuracy level between 79.05% and 74.6% and an estimated increase in profits of between 116.74% and 253.52%. Taking into consideration the declining accuracy level, using a threshold of 0.60 would result in an estimated accuracy level of 78.1% and an estimated increase in profits of 226.81%.