Predicting Loan Defaults with Logistic Regression

Matt Allen

06/26/2018

## 1. Executive Summary

## 2. Introduction

The purpose of this project is to build a model to predict whether or not a borrower will pay back a loan. The desciption of the data set is [here](https://datascienceuwl.github.io/Project2018/loans50k.csv). The model will give a simple “Good” or “Bad” indicating the quality of the loan.

## 3. Preparing and Cleaning the Data

The first step in preparing the data is to create a new column named repsonse that will have the values “Good” and “Bad”. It is based on the variable status. Only the rows with status of “Fully Paid”, “Charged off”, and “Default” will be kept.“Fully Paid” maps to “Good”, and “Charged off” and “Default” are mapped to “Bad”.

The initial dataset consisted of 50,000 records. After reducing “Good” and “Bad”, the total number of records left were 34,655 records. The number of records with the variable response equal to “Good” is 27,074, and “Bad” is 7,581.

There are columns that can be removed, because they are not useful. For example, the status variable may be removed, because it has been transformed into the response variable. The totalPaid variable can be removed, because it is not knowable at the time of issue of a loan.

The variable employment, which indicates job title, can be removed because there is so much variation. It would be possible to create a new variable with job title that puts titles in more general categories, but it probably is still not useful. A better indicator may be income from job. The variable loanID can be removed. It is not necessary in the loan analysis, and may have ethical implications as it has the potential to trace a loan back to an individual. The variable verified indicates verification of annual income completed. This can be removed, because the current study is not meant to test the loan applicaitons truthfulness.

Some of the categories like medical and debt\_consolidation if used in the model could introduce bias against particular vulnerable groups that despite some setbacks may still be able to pay off a loan. Although the variable state is broader than zip code, it can still be discriminatory by geography. Also since state is a somewhat general geography, it may not have much predictive power, and can be removed. The variable length indicates the length of employment. This could be a good indicator, but there are more than 1300 missing values. It should be removed, because removing the rows with missing income would significantly reduce the data set, and imputation methods may be challenging for this much missing data where there is not necessarily a good proxy.

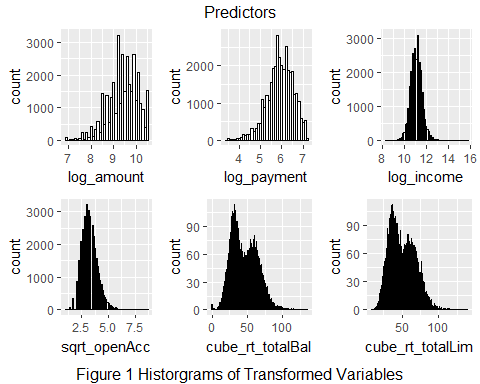
The variable revolRatio indicates proportion of revolving credit in use. This can be removed, because it is captured in other variables like debtIncRatio. With similar resoning, bcRatio can be removed. totalBal and totalLim can be used as proxies for totalRevLim, totalRevBal, totalBcLim,bcOpen, and totalIlLim. Further debtIncRatio, avgBal can be removed because they can be derived from other variables like income, totalBal, and openAcc. Remove totalAcc, because we are interested in accounts that are currently open and this is already captured in openAcc. Remove the variable term, because payment captures this information. The loan term and payment amount are dependent on each other. The variable accOpen24 can be removed, because if an account is opened there will likely be a credit inquiry, which is captured in inq6mth.

The variable grade indicates the risk of the loan. I would consider it to be another response variable like status or the response variable derived from status, but with more levels. For this reason, it should be eliminated. The variable rate can be removed for similar reasons as grade. The loan rate is an indication of the borrower’s risk. The greater the risk the higher the rate. The variable pubRec can be eliminated, because the variable delinq2yr should captures this information. If someone has pubRec against them, they would have also missed payments.

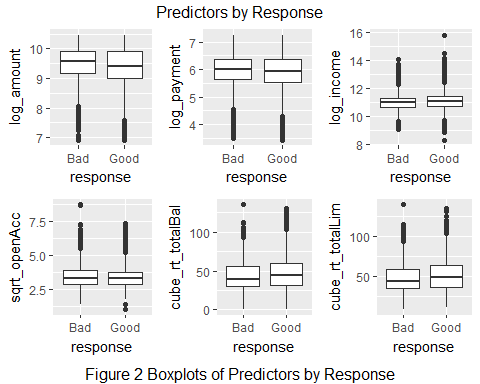
After elimination of variables, the predictors that are left are amount, payment, home, income, delinq2yr, inq6mth, openAcc, totalBal, and totalLim. I converted income and payment to numeric, because they were showing up as factors. With the remaining vairables, there is no imputation or record removal needed.

## 4. Exploring and Transforming the Data

The variables amount, payment, income, openAcc, totalBal, and totalLim showed varying degrees of right skewness. From inspection of histograms, the variables amount, payment, and income showed the most right skewness and were transfored by the log function. The variables totalBal and totalLim were transformed by taking the cubed root. The variable openAcc was transformed by the sqrt function. Historgrams of the transformed variables are displayed in figure 1.



To start to develop a picture of the relationships between the predictor variables, and the response, two boxplots were created for a predictor variable. The boxplots were created to explore the differences in distributions between “Good” and “Bad” loans for a predictor variable. The plots can be seen in figure 2. Looking at the log\_amount variable box plot, the mean of amount is lower for Good than Bad, which makes sense. A smaller loan would be easier to pay back. Similarly with the variable log\_payment, a loan that has a smaller payment makes it easier to make the payment each month. Also as indicated by log\_income, those with greater incomes have more money to pay back a loan.



## 5. The Logistic Model

## 6. Optimizing the Threshold for Accuracy

## 7. Optimizing the Threshold for Profit

## 8. Results Summary