Predicting Loan Defaults with Logistic Regression

Matt Allen

07/29/2018

## 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 description 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. The project will begin by preparing and cleaning the data. Variables will be selected by domain knowledge versus an automated method. Through exploration of the selected variables, the variables may be transformed to satisfy conditions of model fitting like being normally distributed for example. The data will be fitted to a logistic model with a training data set. Part of the data set will be held out as test data set to validate the model. The model is binomial and the mean also called the classification threshold can be adjusted to optimize for accuracy or profit.

## 3. Preparing and Cleaning the Data

The first step in preparing the data is to create a new column named response 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 variables 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 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. Also, amount can be removed, because it is related to payment.

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 capture this information. If someone has pubRec against them, they would have also missed payments.

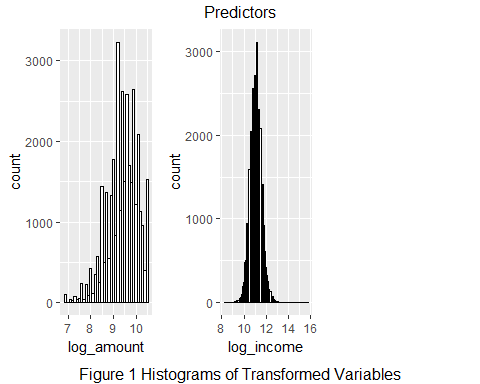
After inspection of the histograms of the variables totalBal and totalLim, it was discovered that there distributions are bimodal, which violates condition of normality, so debtIncRatio was used as a proxy. The variable payment was found to be strongly correlated with amount, and was removed. The variable openAcc was found not to be significant and was removed. The variable loanid was removed because it is just an identifier and has no predictive power.

delinq2yr and inq6mth are highly right skewed

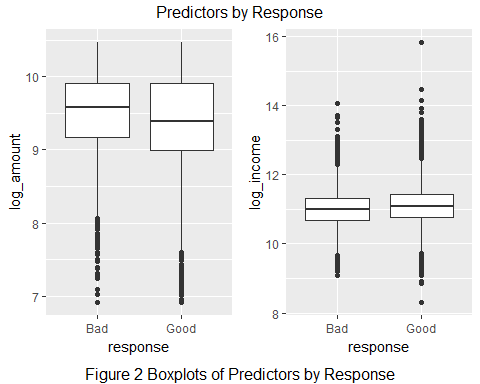
After elimination of variables, the predictors that are left are amount, home, income, debtIncRat, delinq2yr and inq6mth. I converted income and amount to numeric, because they were showing up as factors. With the remaining variables, there is no imputation or record removal needed. The variable home is categorical, the rest are quantitative.

## 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 transformed 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. Histograms 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.



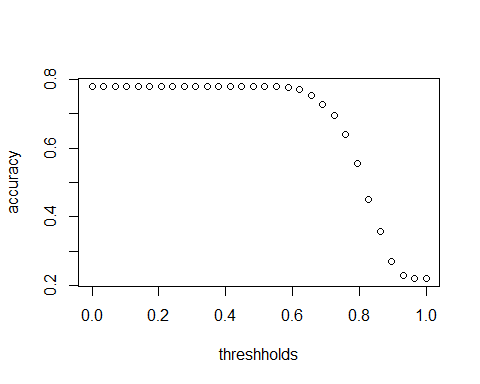
## 5. The Logistic Model

## predLoan  
## Bad Good Sum  
## Bad 6 1521 1527  
## Good 5 5399 5404  
## Sum 11 6920 6931

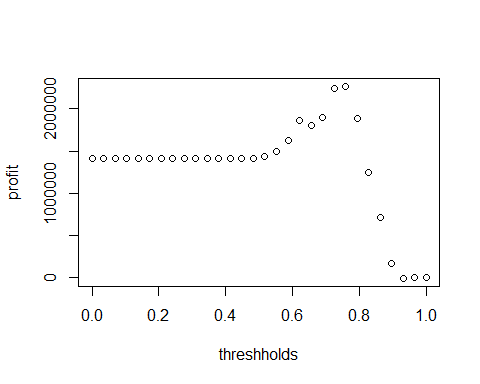
## [1] "Proportion correctly predicted = 0.779829750396768"

## 6. Optimizing the Threshold for Accuracy

## [1] 0.7796855 0.7796855 0.7796855 0.7796855 0.7796855 0.7796855 0.7796855  
## [8] 0.7796855 0.7796855 0.7796855 0.7796855 0.7796855 0.7796855 0.7795412  
## [15] 0.7793969 0.7795412 0.7782427 0.7756456 0.7703073 0.7541480 0.7268792  
## [22] 0.6942721 0.6397345 0.5564854 0.4507286 0.3559371 0.2706680 0.2288270  
## [29] 0.2204588 0.2203145

 ## 7. Optimizing the Threshold for Profit

## [1] 1414003.0143 1414003.0143 1414003.0143 1414003.0143 1414003.0143  
## [6] 1414003.0143 1414003.0143 1414003.0143 1414003.0143 1414003.0143  
## [11] 1414003.0143 1414003.0143 1414003.0143 1407659.1568 1412851.7087  
## [16] 1437544.6476 1495318.5445 1630655.6369 1866902.5305 1808234.6704  
## [21] 1899977.6480 2246771.3516 2267091.6046 1888458.9946 1251243.1485  
## [26] 709985.0068 172995.0737 -8003.6832 631.5367 0.0000



## 8. Results Summary