Title: Variation of interest rate for a constant FICO score

Introduction:

"An interest rate is the rate at which interest is paid by a borrower (debtor) for the use of money that they borrow from a lender (creditor). Specifically, the interest rate is a percent of principal paid a certain amount of times per period (usually quoted per annum)" [1]. The Lending Club [2] offers loans to individuals and determines the interest rate for a borrower by considering several factors like credit score, loan amount, loan term, open credit lines and credit history [3].

The purpose of this analysis is to identify and quantify associations between the interest rate of the loan and the other variables in the data set. In particular, whether any of these variables have an important association with interest rate after taking into account the applicant's FICO score. The results of the anaysis suggest that higher interest rate is associated with increased funded amount and loan length.

Methods:

Data Collection

This analysis used a sample of 2500 loans from the Lending Club, as provided by instructor Jeff Leek for the Data Analysis class on Coursera [4]. These data were downloaded from the course website on November 17th, 2013 using the R programming language [5].

Exploratory Analysis

Exploratory analysis was performed by examining tables and plots of the observed data. Transformations to perform on the raw data were identified on the basis of plots and knowledge of the scale of measured variables. Exploratory analysis was used to:

- identify missing values
- convert labelled variables to their numerical equivalents(e.g., change 60 months to 60 in case of loan length, change <1 year to 0 and 10+ years to 10 in case of employment length)
- survey plots and tables to look for patterns and correlations
- determine the terms used in the regression model relating interest rate to FICO score based on Singular Value Decomposition (SVD) [6]

Statistical Modeling

To relate interest rate to FICO score a standard multivariate linear regression model [7] [8] was used. Model selection was performed on the basis of exploratory analysis. Factors for the model were selected based on the outcomes of the SVD and co-efficients were estimated with ordinary least squares [9].

Results:

The exploratory analysis identified some missing or erroneous values, but these observations were retained for the linear model as the missing values did not appear in the factors under consideration. Some outliers were also found, but these were also retained as there seemed to be effectively no difference in the coefficients of the linear regression model whether they were included or not.

The factors under consideration were identified based on the values of the SVD. The values of SVD suggested that only factors accounting for greater than 10% of the variance should be kept. As such, the factors retained for further study were:

- Amount Requested (24.04%)
- Amount Funded (12.29%)

- Loan Length (10.57%)

Early analyses suggested a relationship between applicant FICO scores and the interest rates(IR) of the loan (Figure 1). As suggested by Figure 1, lower FICO scores emerge as an indicator of higher interest rates. However, given the parameters of the analysis [4], interest rates are the outcome measure and applicant FICO scores are held as constant.

To inspect the factors identified by SVD, additional plots were used to analyze the relationships between those factors, the outcome (interest rate), and the constant (FICO score). Replotting of the FICO score and interest rate data were done, coloring the points by amount requested (Figure 2), amount funded (Figure 3), and by the loan length(Figure 4).

The final regression model that looked at the interest rate (IR) as the outcome and examined the amount requested (AR), amount funded (AF), and the loan length (LL) was:

$$IR = b_0 + b_1((AR) + (AF) + (LL)) + e$$

where b_0 is an intercept term and b_1 represents the change in interest rate associated with the identified factors: loan amount requested, the actual funded by investors, and the length of the loan. The regression model includes an error term (e) to represent all of the unmeasured and unmodeled sources of variance in the interest rate.

A highly statistically significant relationship between interest rate and loan length ($P < 0.001$) and a statistically significant relationship between interest rate and amount funded ($P = 0.003$) was found. No significant relationship was found between interest rate and amount requested. A change in the loan length corresponded to a change of $b_1 = 0.14$ in the interest rate (95% Confidence Interval: 0.13, 0.16) whereas a change of one unit (i.e. $1000) in the amount funded corresponded to a change of $b_1 = 0.0117$ in the interest rate (95% Confidence Interval: <0.0001, 0.0002).

Conclusions:

These analysis suggest that for the same applicant FICO score, a difference in interest rate between two loans can most likely be explained by the length of the loans and the amount funded by investors. Of these two significant factors, the loan length seems to have the greatest effect.

Despite the strong effect indicated by the linear regression model, there remains a possibility that other factors may strongly influence the interest rate of a given applicant's loan. For instance, the SVD pointed to the amount requested as accounting for the most variance in the data, and this implied that the amount requested would also have the strongest effect in the final model. However, the amount requested ultimately did not have a statistically significant effect on the interest rate. This outcome leads to the suspection that other variables from the data, which were otherwise eliminated by the SVD, may affect the interest rate in important ways. Potential confounder variables that may affect the interest rates, for e.g. inflation rate, terms of service, etc, were beyond the scope of this analysis.

Future analyses should probe other factors in more detail to see if the effects and patterns indicated in the linear regression model persist. While this analysis was done on a limited set of data, a larger collection of representative interest loans may be more appropriate for understanding the relationship between various parameters.

References

1. Wikipedia “Interest rate” page URL: http://en.wikipedia.org/wiki/Interest_rate Accessed November 17th, 2013
2. The Lending Club home page URL: https://www.lendingclub.com/ Accessed Novemeber 17th, 2013
3. The Lending Club personal loan page URL: https://www.lendingclub.com/public/personal-loans.action Accessed November 17th 2013
4. Coursera.org: Data Analysis Assignment 1. URL: https://class.coursera.org/dataanalysis-002/human_grading/view/courses/971332/assessments/4/submissions Accessed November 17th 2013
5. R Core Team (2013) "The R Project for Statistical Computing" URL: http://www.r-project.org/ Accessed November 17th 2013
6. Singular Value Decomposition Tutorial URL: http://www.ling.ohio-state.edu/~kbaker/pubs/Singular_Value_Decomposition_Tutorial.pdf Accessed November 17th, 2013
7. Multivariate Linear Regression URL: http://www.public.iastate.edu/~maitra/stat501/lectures/MultivariateRegression.pdf Accessed November 17th, 2013
8. Multivariate Linear Regression Models URL: http://www.math.ust.hk/~makchen/Math347/Chap7.pdf Accessed November 17th, 2013
9. Ordinary Least Squares URL: https://datajobs.com/data-science-repo/OLS-Regression-%5BGD-Hutcheson%5D.pdf Accessed November 17th, 2013