Variability in interest rates is associated with FICO scores, loan term and amount, applicant's balance, and frequency of loan requests

Introduction

Interest rates for loans are determined by taking into account various factors, most generally defined as a good credit history and low financial risk of an applicant. According to the Lending Club [1], interest rates (ranging from approximately 6% to 26%) are defined in accordance with a loan grade ranging from A (the highest) to G (the lowest). FICO score is a credit score [2], and it is one of the defining factors, along with other factors defining an applicant's credibility. However, data from the Lending Club indicates that while there is an association between higher FICO score and lower interest rate, there are cases when the same interest rate is assigned to individuals with different FICO scores (sometimes the difference can be as great as 100 points).

It is thus evident that the FICO score alone is not sufficient to explain variation in interest rates. Understanding what other variables are at play in determining interest rates is vital for individuals wishing to qualify for the lowest rate possible. Here an analysis was performed in order to determine what other factors are associated with changes in interest rates. An exploratory analysis and a multiple regression analysis reveal that there are additional factors (loan term and amount, applicant's loan balance and frequency of loan requests) responsible for determining an interest rate.

Methods

Data Collection

Data used for this analysis comes from the Lending Club at https://www.lendingclub.com/home.action. Data was downloaded on Wednesday, November 6th 2013, from https://dl.dropbox.com/u/7710864/courseraPublic/exampleProject.zip - a link provided on Coursera's Data Analysis class webpage. Data was downloaded with R programming language [3].

Exploratory Analysis

Exploratory analysis was performed in order to identify the quality of data (such as identifying and excluding missing values) and to determine the proper terms for the regression model in order to model a relationship between interest rate and other variables. Exploratory analysis consisted of a) transforming the raw data (e.g., identifying missing values, performing log transformations and converting variables into numeric and factor formats, as needed); and b) studying data plots.

Statistical Modeling

In order to identify the relationships between interest rates, FICO scores and other variables of interest, a multivariate linear regression model [4] was performed. The choice of the model was informed by the results of the exploratory analysis, a series of simple linear regression models as well as the educated guesses on the nature of how interest rates are determined.

Reproducibility

All analyses have been reproduced in the R markdown file loansCode.rmd [5], which can be provided upon request. Data used for analysis can be provided at the same time as it is not available on the Lending Club's website.

Results

Dataset used in this analysis contains 2500 observations and 14 variables: awarded interest rate (Interest.Rate), requested amount of money (Amount.Requested), purpose of the loan (Loan.Purpose), monthly income (Monthy.Income), how many times a loan was requested in the last 6 months (Inquieries.in.the.Last.6.Months), how much money was awarded (Amount.Funded.By.Investors), debt to income ratio (Debt.To.Income.Ratio), length of the employment (Employment.Length), goodness of the credit score (FICO.Range), state of residence (State), how many other credit lines are open (Open.CREDIT.Lines), when the loan will be repaid (Loan.Length), home ownership status (Home.Ownership), and how much balance is left on the loan (Revolving.CREDIT.Balance).

Dataset contained 7 missing values, which were excluded from the analysis. Additionally, 8 more values were classified as missing and they were excluded as well. Thus, there were two negative values (-0.01) in a column containing information on how much money was awarded. There were also six "n/a" values in a column containing information on the length of employment. Considering that the number of these values represented only 0.32% of the total dataset and there was no conclusive explanation as to the meaning of these values, it was decided to exclude them from the analysis.

Additional manipulation of data included converting the credit score from a range (represented as a factor data type) to a minimum numeric value of a corresponding range(FICO); and converting such factor variables as length of employment, interest rate, debt to income ratio, and length of loan term into numeric variables as well. Variables related to dollar amounts (i.e., monthly income, requested amount of money, and how much money was awarded) have larger magnitude of values relative to other numeric variables, and in order to bring all numeric values to a similar scale, a log10 transformation was performed. Log transformation also helped to reduce any skewedness of data. Discussion of all subsequent analyses involves transformed variables only.

A simple linear regression model relating interest rate to FICO scores explains about 50% of the variation (adjusted R^2 = 0.5011). In order to improve the model and to determine which factors explain the rest of the variation, a series of simple linear regression models relating interest rate to other variables were undertaken. The analysis reveals that such variables as revolving balance, amount of money requested, length of the loan, debt ratio, and money awarded are related to interest rate at a confidence level of 95-99%, as indicated by corresponding p-values. However, only length of the loan term explains 18% of the variation in the model (adjusted R^2 = 0.18), while others explain less than one per cent on average.

A final regression model took the following form:

$$IR = b_0 + b_1(FICO) + b_2(loanLength) + b_3(numInquiries) + b_4(revolvingBalance) + b_5(amountRequested) + e,$$

where b_0 is an intercept term and b_1 represents one percent change in interest rate associated with a change of one unit of *FICO* scores for all applicants who have the same loan term (*loanLength*), number of inquiries (*numInquiries*), revolving balance (*revolvingBalance*) and amount of money requested (*amountRequested*). The error term *e* represents all sources of unexplained and unmeasured random variation in interest rates.

There is a highly significant statistical association ($p < 2.2e-16$) between interest rates and FICO scores. An increase of 5 units in FICO score decreased interest rate by 0.088% (95% CI: -0.09, -0.85). Thus, keeping all other variables the same, the interest rate is expected to differ by 0.44 units for each pair of adjacent FICO scores in increments of five (e.g., 640 vs. 645; 645 vs. 650; 720 vs. 725 etc.).

Conclusions

This analysis suggests that there is a statistically significant association between interest rates and such variables as FICO scores, loan term, number of inquiries in the last six months; amount requested by the applicant and applicant's revolving balance. While FICO scores and the amount of revolving balance are associated with decreased interest rates, other variables listed above contribute to higher interest rates. It must be noted that since two variables – revolving balance and amount of money requested – have been log-transformed, their estimated effects are not linear. Inclusion of such variables as loan term, number of inquiries in the last six months, amount requested by the applicant, and applicant's revolving balance improved the adjusted R^2 of the model from 50% to 75% of explained variation (**Figure 1**).

There are several limitations of the model. While confounders show a statistically significant association with interest rate, the effects appear to be very small (often representing as little as less than one percent change in interest rates). Adding and removing other variables as confounders frequently changed significance levels, suggesting that there are other interactions in the model that have not been fully explored by this analysis. Finally, this analysis only concentrated on numeric variables and did not take into account such factor variables as home ownership and purpose of the loan. However, the purpose of this analysis was to investigate what variables other than FICO score are associated with changes in interest rates. As this was the first step in tacking the issue, further investigation will be necessary in order to explore more fully what factors have predictive effects on interest rates. It is possible that additional factors not represented in the dataset (e.g., criminal history of an applicant) should be taken into account.

References

1. Lending Club. URL: https://www.lendingclub.com. Accessed 11/12/2013
2. myFICO. URL: http://www.myfico.com/crediteducation/whatsinyourscore.aspx. Accessed 11/13/2013
3. R Core Team (2012). "R: A language and environment for statistical computing". URL: http://www.R-project.org
4. Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 936. Wiley, 2012.
5. R markdown Page. URL: http://www.rstudio.com/ide/docs/authoring/using_markdown. Accessed 11/12/2013