Relationships between Interest Rates and Characteristics of Applicants in Personal Loans

Introduction

Personal loan, also known as unsecured loan, is a type of loan that requires no collaterals and usually involves a small amount (< $35,000) for purposes of computer purchases, home improvements, vacations, education expenses, etc [1,2]. Because there are no assets pledged in the lending agreement to enforce the repayment, such personal loans are considered riskier and therefore have stricter approval standard and charge higher interest rates. Moreover, the interest rate is determined on a case-by-case basis, highly depending on the applicant's credit information and other characteristics. The most important factor to consider is the credit score, which is a measure of a person's creditworthiness. In the US, the most widely utilized credit score model is the FICO score. It is a compound metric that already captures various personal financial factors such as late payment rate, the length of credit history, credit utilization ratio, amongst others, and the value ranges from 300 to850 [3].

The exact formula behind the interest rate computation is usually a secret and varies from one loan service provider to another. However, it is very interesting to investigate the relationships hidden in the black box from a general perspective. It helps one gain insight into what factors, alongside with the credit score, affects the determination of the interest rate and how they do it. Here we carry out an analysis to indentify and quantify the association between interest rate and a few personal factors of the applicants. We use exploratory analysis and multiply regression analysis in this study. We are able to find that the interest rate shows a negatively linear relation to the credit score, and it is also significantly associated with factors including the loan length, the amount of the loan, applicant's monthly income, the number of open credit lines, and the number of recent credit queries.

Methods

Data collection

The data used in this study consist of a sample of 2500 personal loans issued by the Lending Club [4], and are available as resource on Cousera for the *Data Analysis* course by Jeff Leek [5]. The data were downloaded on Nov. 11, 2013.

Exploratory Analysis

The exploratory analysis is performed in two stages. In the first stage, we examine the data using tables and plots and calculating correlations to identify missing values and obtain qualitative relations between variables. Two records with multiple missing values are removed from the raw data before proceeding to further statistical analysis. After a basic regression model is built for the interest rate, we perform another exploratory study on the residuals to determine extra terms to include in the model.

Statistical Modeling

We use multivariate linear regression model [6] to relate the interest rate to a few financial factors. Model selection is based on the observation in the exploratory analysis as well as knowledge on the basics of setting interest rate from the Lending Club [7]. Coefficients are estimated with ordinary least squares and standard errors are calculated using standard asymptotic approximations [8].

Results

The loan data contains the following variables: the interest rate (IR), the amount requested (AR), the amount loaned (AL), the loan length (Len), the loan purpose (Pur), the debt-to-income ratio of the applicant (DTIR), the state (Sta), the type of home ownership (HO), the monthly income (MI), the length of current employment (LCE), the range where the applicant's FICO score lies (FICO), the number of open credit lines (NCL), the revolving credit balance (RCB), and the number of recent Inquiries for credit (NRI). Note that the loan length variable is binary, consisting of two values -- short-term (36 months) and long-term (60 months). Also we take the midpoint of each FICO range as the FICO score.

An important observation through the exploratory analysis is the clear relations between the interest rate and the FICO score and the loan length, as shown in t Fig. 1(a). The dataset is divided into two subgroups based on the loan length variable, and the average interest rate of each FICO score range is also computed and shown. We found that the long-term interest rates have greater values on average than the short-term ones. Also, a two-segment linear relation between the interest rate and the FICO score is suggested in the plot. One can see that in the short-term data, for FICO scores greater than 730, the averaged interest rate is significantly more robust to the change of FICO scores. Similarly but less obviously, the long-term data show a critical point at approximately 760. This behavior can be understood as that, for applicants who have high FICO scores above a certain threshold, they are considered to be in similar top tiers in the evaluation process. To include this information, we code into the data a new binary variable indicating whether the FICO score is above the critical point or not (FaC), and use 730 and 760 as the threshold values for the short-term and long-term data, respectively.

To avoid complicated interaction terms that are hard to interpret, we treat the long-term and the short-term loan data independently, are hence exclude the loan length variable in the our modeling. We fit a basic model relating the interest rate to the FICO score and the threshold indicator variable as follows

$$IR = b_0 + b_1\, FICO + b_2\, (FAC=True) + b_3\, FICO * (FaC=True) + e.$$

b_0 is an intercept term and b_1 represents the change of interest rate associated with the unit increase in the FICO score in the range below the critical point. In the range above the critical point, the intercept and the change rate is $(b_0 + b_2)$ and $(b_1 + b_3)$, correspondingly. The error term *e* represents variation that is not explained by the model.

However, the residuals from this basic model exhibit patterns of non-random variation. Figure 1(b) shows the distribution of residuals grouped by different ranges of requested loan size for the short-term data. The observation that larger loans have greater values of residuals on average violates the assumptions of homoscedasticity, indicating that the loan amount is likely to relate to the interest rate. To proceed, we perform a second exploratory analysis on the residuals and identify potential dependent factors to include in the final model. The adjusted model becomes

$$IR = b_0 + b_1\, FICO + b_2\, (FAC=True) + b_3\, FICO * (FaC=True) + b_4\, AR + b_5 \log_{10}(MI) + b_6\, NCL + b_7\, NRI + e.$$

b_4 represents the change of the interest rate with 1-dollar increase in loan amount requested, fixing all other variables. Similar interpretation extends to b_5, b_6, and b_7. The inclusion of the confounders addresses the non-random variation issue presented in the basic model. The residuals distribution of our full model is shown in Fig. 1(c), where one sees the loan size related pattern is eliminated.

We report that the interest rate is significantly related to all the model variables for the short-term loan data ($p < 2.2e\text{-}16$), and so it is for the long-term data ($p < 2.2e\text{-}16$) except for the number of open credit lines. For short-term loans, an increase of 1 in FICO score (other variables fixed) results in a decrease of $b_1 = 0.00115$ (95% Confidence Interval: 0.00110, 0.00120) in the interest rate if the FICO score is below 730, while the number becomes $(b_1 + b_3) = 0.00026$ (an order of magnitude smaller) if the FICO score is above the critical point. An increase of $1 in the amount requested results in an increase of b_4 =1.53e-6 (95% Confidence Interval: 1.39e-6, 1.67e-6) in the interest rate; 1 extra open lines for credit results in an increase of $b_6 = 0.00355$ (95% Confidence Interval: 0.00287, 0.00423) in the interest rate. Therefore, the same applicant is expected to face an interest rate 0.015 greater if he requests a short-term loan $10,000 higher. Similarly, for two applicants with the same conditions except for the number of open credit lines, the one with 5 lines is expected to be offered an interest rate 0.00355 higher than the other person with 4 lines.

Conclusions

Our analysis suggests that the interest rate of personal loans is significantly associated with the FICO score, the loan length, the amount of the loan, monthly income, the number of open credit lines, and the number of recent credit inquiries. In addition, being above the FICO score critical point considerably decreases the influence of FICO score on the interest rate. We estimate the relations using a multivariate linear regression model.

Meanwhile, we expect that additional variables may contribute to a better model. For example, instead of the revolving credit balance of the applicants, the credit utilization ratio (balance/total credit available) is a better factor to consider. Also, the knowledge that most of the factors affect the interest rate not on a simple linear basis but on a non-uniform region basis

[7], just like the FICO score, suggest that a decision tree model may help to identify regions and reveal the non-linear relations between the interest rate and the dependent factors.

References

1. LendingTree.com “Unsecured loans - risky business for lenders” Page. URL: https://www.lendingtree.com/personal-loan-basics-summary. Accessed 11/16/2013.
2. Wikipedia “Unsecured Debt” Page. URL: http://en.wikipedia.org/wiki/Unsecured_debt. Accessed 11/16/2013.
3. Wikipedia “Credit Score in the United States” Page. URL: http://en.wikipedia.org/wiki/FICO_score#FICO_score. Accessed 11/16/2013.
4. Lending Club Home Page. URL: https://www.lendingclub.com/. Accessed 11/11/2013.
5. Resource Like Provided on Coursera – Data Analysis Website. https://spark-public.s3.amazonaws.com/dataanalysis/loansData.csv. Accessed 11/11/2013.
6. Seber, George AF, and Alan J. Lee. Linear regression analysis. Vol. 936. Wiley, 2012.
7. Lending Club “Interest Rates and How We Set Them” Page. URL: https://www.lendingclub.com/public/how-we-set-interest-rates.action. Accessed 11/17/2013.
8. Ferguson, Thomas S. A Course in Large Sample Theory: Texts in Statistical Science. Vol. 38. Chapman & Hall/CRC, 1996.